

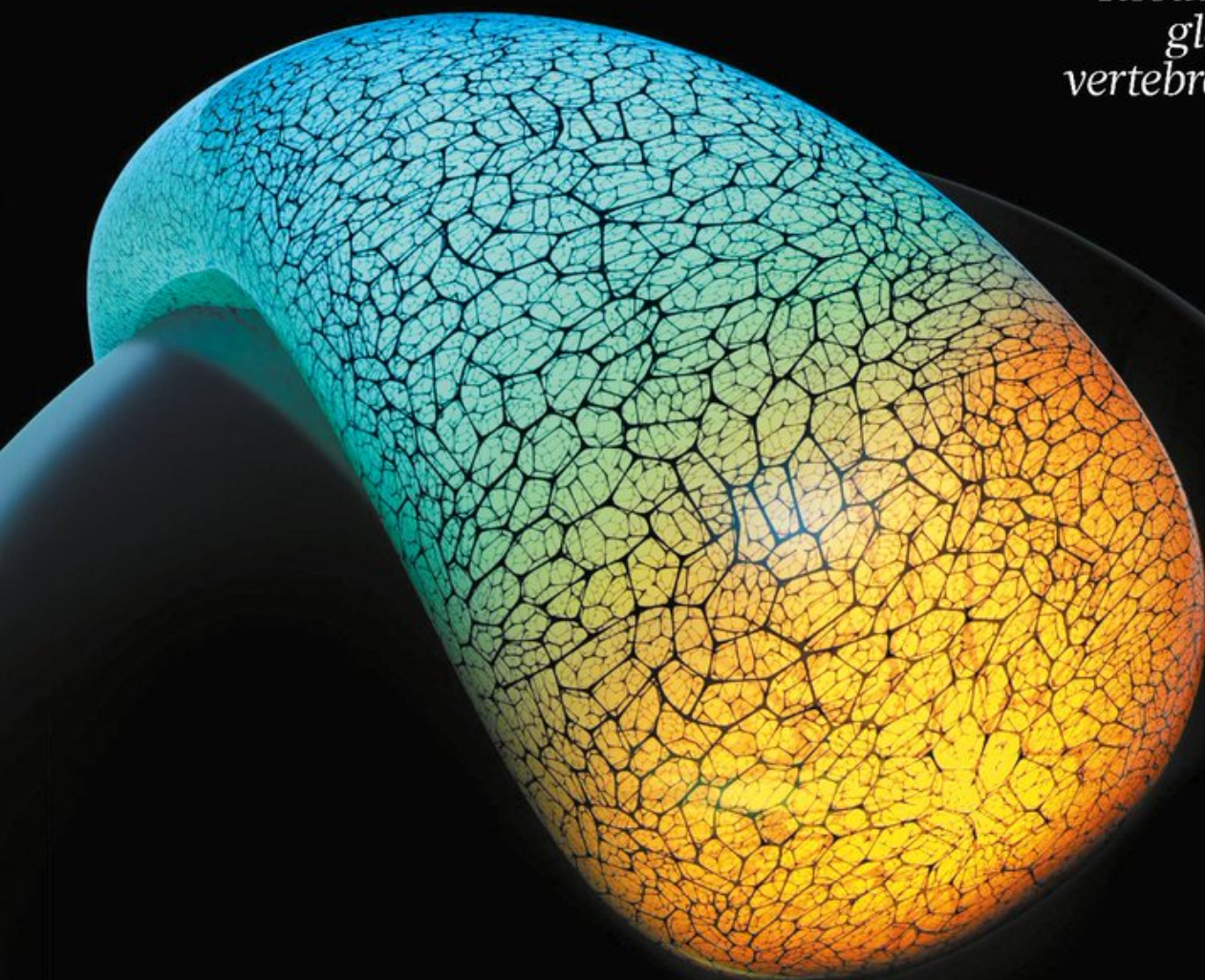
nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

GROW WITH THE FLOW

*Tissues 'melt' like
glass to shape
vertebrate embryos*

PAGES 315 & 401



CLIMATE CHANGE

THE ONLY WAY IS ETHICS

*Social effects of carbon
capture should be considered*

PAGE 303

HIGH-ENERGY PHYSICS

PARTICLE ACCELERATION

*Electrons surf a proton-
driven plasma wakefield*

PAGES 318 & 363

STRUCTURAL BIOLOGY

NEGATIVE RESULT

*Channelrhodopsins offer
route to anionic optogenetics*

PAGES 312, 343 & 349

NATURE.COM

20 September 2018

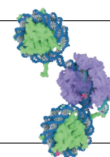
Vol. 561, No. 7723

THIS WEEK

EDITORIALS

CATALYSIS Clever chemistry could convert biomass to jet fuel **p.286**

WORLD VIEW Update undergraduate training to include reproducibility **p.287**



HANDS OFF! Bundles of cellular DNA defy gene editing **p.289**

Something in the air

High levels of air pollution are dangerous, damaging and a global disgrace. Science can help by offering better models for monitoring and exposure.

Air pollution was once celebrated. Industrialists in Victorian Britain would point to the smoky streets of the Industrial Revolution and see only the signs of wealth and progress. Alerted in the 1960s to the stink of an Alabama paper mill some 30 kilometres away that was reaching the state capital, Governor George Wallace remarked: “Yeah, that’s the smell of prosperity.”

Public attitudes have changed. Clean air to breathe is widely recognized by the United Nations and others as a universal human right, essential to physical well-being. But a change in mindset about air does little to actually clean it. More than four million people still die each year from exposure to polluted outside air — an intolerable situation, and one that is perpetuated by urbanization and regulatory impotence.

Nine out of ten people live in places where outdoor air pollution exceeds guidelines set by the World Health Organization (WHO). Hotspots are congested urban areas in low- and middle-income countries such as India, Nigeria and China. In some megacities — Mexico City, for example — authorities have begun to adopt cleaner vehicle standards. But fine particulate matter and nitrogen dioxide from vehicular traffic, energy production, industry and heating remain a serious public-health risk in most built-up areas.

Even many cities in wealthy Europe fail to meet the WHO standards. A report last week by the European Court of Auditors, which regularly scrutinizes the effectiveness of European Union policies and programmes, concludes that action taken so far to improve air quality is not sufficiently protecting citizens from pollution. Cities that auditors visited for the report — including Brussels, Kraków, Milan and Sofia — have made little or no progress since 2009 in reducing particulate matter pollution (Kraków and Sofia) or since 2012 in reducing nitrogen dioxide levels (Brussels and Milan). Although emissions of air pollutants have been decreasing overall, most member states still do not fully comply with stringent EU air-quality standards set up in 2008.

The European Commission has already taken several member states to court over their failure to introduce appropriate measures. Meanwhile, a 2015 scandal over faked Volkswagen vehicle emission tests in the United States has helped to bring the problem to greater public and political attention by offering a corporate villain. Low-emission zones in London (a persistent offender when it comes to breaching clean-air regulations) and many other European cities now ban badly polluting vehicles or restrict their access. That is good news for some metropolitan neighbourhoods, but it is only a first step. Little overall benefit is gained, for example, if diesel cars that are no longer wanted in Europe are pushed by manufacturers into markets abroad.

Effectively tackling the causes and effects of air pollution requires a more joined-up approach. Air-quality regulations in the EU, for example, must be taken into account more fully when setting policies on climate, transport, enterprise, trade and innovation.

Science, too, can do more to mitigate health risks from poor air quality. It is important to unpick how different types and levels of pollution affect human health. The epidemiological research needed

to do that requires more-consistent methodologies to monitor and report pollution and human exposure to it.

Scientists can also help to develop and provide well-tested modelling tools that local authorities can use to improve assessments of their specific circumstances, and to design action plans. The Forum for air quality modelling in Europe (FAIRMODE), a joint programme by scientists with the European Commission’s Joint Research Centre and the European Environment Agency, is tasked with developing air-pollution models and is working to harmonize monitoring methods across the bloc. But as air-quality concerns continue to grow, the forum must liaise more with city leaders and health specialists to make sure they get the tools and data they need.

The results of this environmental science should be shared with countries worldwide. The situation is bad in rich countries: the WHO says that about half of city dwellers in developed nations are exposed to air that does not meet its guidelines. In cities of more than 100,000 people in the developing world, that figure rises to include almost everybody (97%). India alone has nine of the world’s ten most-polluted cities. Air is a shared resource. Research and tools to make it safe to breathe should be shared as well. ■

“This environmental science should be shared with countries worldwide.”

Fighting fraud

An Austrian success story shows one way to tackle misconduct.

Many countries are trying to clamp down on scientific misconduct. Last week, the UK government promised to look into setting up an independent body to oversee institutional investigations into research misconduct, and the Netherlands has revamped its research-integrity code. Last month, India said it would crack down on widespread academic plagiarism. And earlier this year, Chinese officials pledged to get tough on academic fraud with new laws that include a dedicated government agency to police misconduct.

The problem is that much of this renewed political attention is not translating into meaningful action. High-profile cases of exposed malpractice continue to pile up, and surveys of researchers regularly confirm that poor behaviour is shockingly more common than many who promote the values of science might want to accept.

So it is promising to report from a meeting in Vienna last week that was held to celebrate ten years of the Austrian Agency for Research Integrity. The organization is not perfect, but it has much to be proud of. Its work shows what can be achieved given the requisite political

will. And it reveals some of the problems that remain, in Austria and elsewhere. Officials in countries that are looking for ways to tackle misconduct should pay close attention.

Lesson one: act quickly and decisively. The agency was born out of a scandal that rocked Austrian science to its core. In 2008, the Austrian Agency for Health and Food Safety deemed a clinical trial of an experimental therapy for urinary incontinence to be illegal and invalid. The trial, led by Hannes Strasser at the Medical University of Innsbruck, was conducted without appropriate approvals, and did not adequately inform or protect patients. But the university initially failed to investigate.

At the time, an Editorial in *Nature* lamented the sorry state of Austrian science, which was riddled with rigid hierarchies that deterred many from raising complaints and concerns (*Nature* 454, 917–918; 2008). The article called for the nation to speed up the creation of an independent body to investigate cases of academic fraud, which it had been planning and discussing for some time.

It did so. Since June 2009, the agency has handled 144 allegations of research fraud, and confirmed 40 cases. Of the rest, 12 are ongoing. In 31 cases, it was not possible to determine whether misconduct had occurred, and for a further 37 the allegations were not within the remit of the agency (for the most part, these revolved around labour disputes). The remaining 24 were either not followed up or were investigated by the university in question.

Lesson two: institutions have nothing to fear. The Vienna agency offered a confidential route for research scientists to report concerns, but required institutions to buy in to the agency by becoming members. Initially, many universities were reluctant to sign up, fearing their reputations could be ruined if they were found to be harbouring fraudsters. But the ministry of higher education linked membership to funding, which quickly persuaded them to change their minds. All of the country's 22 public universities have now signed up. Sanctions against

researchers found to have committed misconduct are left to the universities. According to the agency, these include sackings and retractions.

Lesson three: one size cannot fit all. Any investigatory system must consider unique aspects of a country's research system. The Austrian agency, for example, uses scientists working outside the country to assess the complaints. This is crucial for protecting the process from undue influence from strong local networks and loyalties within the small nation's academic research community of fewer than 20,000 people.

“Research misconduct is moving higher up the political agenda.”

Lesson four: wider legal reforms are necessary to properly address cases of fraud. Much behaviour that science frowns on is not explicitly against the law, and findings of misconduct and associated penalties can themselves be challenged in court. In 2012, the Austrian agency concluded that protein crystallographer Robert Schwarzenbacher had faked the structure of a birch-pollen allergen. Schwarzenbacher lost his job at the University of Salzburg, but later sued the institution for unfair dismissal. The case was settled out of court. In 2011, an employment tribunal ordered that Hannes Strasser be readmitted to a teaching post at the Medical University of Innsbruck. (He lost that post in 2014 when a final criminal-court ruling sentenced him to jail for aggravated libel related to the case.)

The legal status of scientific fraud is a thorny issue — and one hotly debated. But Sweden, following Denmark, is already working to define research misconduct in law so that there are clear lines in place. Laws against misconduct would also compel more institutions, such as those that are privately funded, to act transparently.

Research misconduct is moving higher up the political agenda. And for countries that are in the process of creating systems, revamping old ones or assessing their achievements, Austria offers a good example to follow. Institutions that continue to drag their feet on the problem should take careful note, too. ■

False fuels

Clever chemistry brings synthetic kerosene and petrol closer.

Necessity is the mother of invention, and a century ago, nations needed petroleum. They could run ships on coal, but burning solid lumps of fuel was impractical for cars and tanks, and unsuited to aircraft. Unlike other countries, Germany had no access to crude oil, so two chemists there — Franz Fischer and Hans Tropsch — invented a way to make synthetic petroleum from coal in 1925.

Their Fischer–Tropsch (FT) process could now help countries and companies that want to phase out fossil fuels: if coal can be turned into liquid fuels, then, theoretically, greener alternatives such as biomass could be as well. But so far, efforts to do this have been inefficient, and certainly not cheap enough to compete with oil.

A study in *Nature Catalysis* this week points to a possible way forward. Chemists in Japan and China have boosted the FT process, and improved on how it can be steered to produce different liquid fuels (J. Li *et al.* *Nature Catal.* <http://doi.org/ctxv>; 2018).

Although the FT process is good at converting gases — used directly, or produced from solids such as coal or even ground-up peanut shells — it's rather unfussy about what it churns out. Mostly, that's a blend of synthetic-petroleum products, from light gases such as methane through to heavy waxes (think Vaseline). The most useful stuff, such as petrol, diesel and aviation fuel (kerosene), falls somewhere in the middle, and must be separated and purified. That typically makes large-scale FT synthesis of those fuels a two-step process, which increases costs, complexity and pollution.

As a consequence, it's usually used commercially to make synthetic liquid fuels only where the feedstock is unusually cheap (China operates some facilities that process coal), or where there is no alternative (the South African company Sasol developed an FT process to liquefy coal when access to foreign oil was denied by sanctions in the apartheid era).

The latest study shows that this conversion can be made more selective. With small tweaks to the composition of the catalyst used — a well-known porous material, called a zeolite, mixed with cobalt nanoparticles — the team steered the chemical reaction to produce significant quantities of the desired liquid fuel. For example, the chemists could tune it to make 74% pure petrol (gasoline) or 72% pure jet fuel. Conventionally, it was difficult to produce anything more than 50% using FT synthesis, in a process usually based on iron or cobalt catalysts supported on silica or aluminium oxide. This is one of a string of recent results to show that barrier can be overcome.

There remains some way to go. Zeolite-based catalysts are notorious for their fast deactivation, and the paper reports the synthesis of the fuels in a reactor the size of a thimble, using just a single gram of catalyst. To make it economical, the process would need to be run stably for much longer and scaled up to much larger reactors using at least 100 tonnes of catalyst. Enthusiasm for synthetic fuels ebbs and flows with the market: they were popular a decade ago when oil prices were at record levels, but not so much now. There is no guarantee that the market demand for these fuels will drive the necessary investment.

Noritatsu Tsubaki, a chemist at the University of Toyama in Japan who led the project, says a major advantage of the process is that it could be used to make 'one-step' direct synthesis of kerosene and petrol from FT reactions for the first time — with yields high enough to avoid needing the separation step. Several airlines are already looking into FT chemistry as a source of fuel, and Tsubaki says his team plans to contact airlines and aircraft manufacturers with the findings. The necessity is clearly there, and now, so is a possible invention. ■



Reboot undergraduate courses for reproducibility

Collaboration across institutes can train students in open, team science, which better prepares them for challenges to come, says Katherine Button.

Three years ago, as I prepared to start as a lecturer in the University of Bath's psychology department, I reflected on my own undergraduate training. What should I emulate? What would I like to improve? The 'reproducibility crisis' was in full swing. Many of the standard research practices I had been taught were now shown to be flawed, from *P*-value hacking to 'HARKing' — hypothesizing after the results are known — and an over-reliance on underpowered studies (that is, drawing oversized conclusions from undersized samples).

It struck me that the research dissertation students do in their final year is almost a bootcamp for instilling these bad habits. Vast numbers of projects, limited time and resources, small sample sizes, the potential for undisclosed analytic flexibility (*P*-hacking) and a premium on novelty: together, a recipe for irreproducible results.

Most undergraduate dissertations turn into exercises tallying the limitations of the research design — frustrating for both student and supervisor. However, each year a few students get lucky and publish, securing a huge CV advantage. I wondered what lesson this was teaching. Were we embedding a culture that rewards chance results over robust methods?

In an effort to disrupt this culture, I set up the GW4 Undergraduate Psychology Consortium with colleagues at the universities of Bath, Bristol, Cardiff and Exeter. We wanted to embed rigorous research practices into undergraduate education, incorporating procedures such as pre-registration of study protocols, designing studies with sufficient statistical power and transparent reporting of methods and results.

The difficulty was working out how. Rigorous research methods often take more time and resources than a student project allows. Our solution was collaboration. By working together, students could pool their efforts in data collection to reach sample sizes sufficient for meaningful analyses.

The Consortium is now entering its third year. We are still evolving, but we have settled into a productive routine. It works best if a PhD student or postdoc develops the primary research question for the undergraduates to tackle, drafts a 'bare-bones' study protocol and manages the study. Over the UK summer break, this protocol is circulated to undergraduate students (usually from two to five students at each institution), and each of them plans a secondary research question and suitable method.

At the start of the undergraduates' final year (in the first week of October), we hold the first consortium meeting, where students pitch their secondary questions and decide which will make it into the study. For example, if the main study question is on the effect of impulse-control training on reducing unhealthy food choices, an undergraduate might propose investigating whether effects are moderated by personality traits such as impulsiveness. The student will then propose

a measure for assessing that trait, and propose an analysis to test their hypothesis. This way, each student has some design input, but the sample size and research integrity of the main project is retained. In addition, each student can focus on a slightly different question and so meet requirements for individual assessment. The study protocol is publicly preregistered (in our case, at the Open Science Framework at <https://osf.io>), and data collection runs for four months, from November to March.

In April, students present their findings to the group and collectively discuss the main study results. They reach consensus on conclusions and write up results for wider dissemination.

There are costs. Consortium studies take more time to set up and more effort to coordinate than does the standard student project. But these costs are a small price to pay for giving students the opportunity to network with peers and with researchers at other institutions, exposure to better practices and the feeling of being a valued part of a team. We academics benefit from aligning our teaching with our practice.

It is an example of how, with a bit of creative thinking, we can overcome some of the pitfalls of the current model when it comes to training the next generation to do quantitative experimental research. A handful of publications are in the works.

Both the open-science movement and the growth in online platforms for behavioural tasks and questionnaires have made it easier for psychologists to work across institutions. Using these, we can be confident we are running the same experimental procedures across sites.

Clearly, this approach is not appropriate for all types of research. It might be harder for wet-lab studies, say, in which consumables are expensive, and the idiosyncratic set up of labs makes it more challenging to standardize operating procedures. Yet working collaboratively might be even more beneficial when establishing generalizability or harmonizing methods are more difficult, especially given that students who enter graduate school can sometimes spend years trying to reproduce published work before building upon it.

Early training in collaboration might also bring comfort and creativity with regards to similar approaches later in students' research careers. Although real-world research is increasingly collaborative, it lacks conventions on how to adequately recognize and reward individuals' research input. Perhaps there are wider lessons to take from how we've designed our approach to align rigorous consortium research methods with university requirements for individual assessment. ■

Katherine Button is a lecturer in the department of psychology at the University of Bath, UK.
e-mail: k.s.button@bath.ac.uk

WERE WE
EMBEDDING
A CULTURE THAT
REWARDS
CHANCE
RESULTS
OVER ROBUST
METHODS?

SEVEN DAYS

The news in brief

PEOPLE

Charges dismissed

A judge in Los Angeles County, California, has dismissed a criminal case against chemist Patrick Harran, who faced charges of violating health and safety standards after an accidental death of a young researcher in his laboratory in 2009. The charges stemmed from an incident in which Sheharbano Sangji, a research assistant in Harran's lab at the University of California, Los Angeles, died from third-degree burns incurred during a chemical fire. Sangji was handling *t*-butyl lithium with a syringe when the compound exploded into flames. Under a 2014 agreement with law-enforcement officials, Harran had to meet certain terms — including speaking to university students on lab safety — or see his case go to trial. The judge determined that Harran had met the terms of the agreement, and dismissed the charges on 6 September, according to Thomas O'Brien, Harran's lawyer.

RESEARCH

Fungal alert

Fungal pathogens are spreading rapidly across the world and increasing the risk of diseases in plants and natural ecosystems, warns a report released on 12 September. The report, by London's Royal Botanic Gardens, Kew, is the first to examine the state of fungi worldwide as well as their role in natural ecosystems and human health. Climate change is already affecting the health and reproduction of the organisms, says the report. Fungi first appeared on Earth about 1 billion years ago, and most plants rely on them to thrive. But the kingdom is less well studied than plants and animals because the organisms are perishable and are often

hidden from view. Scientists think that Earth hosts up to 3.8 million fungal species, but only about 144,000 have been classified, at a rate of around 2,000 a year.

AWARDS

Lasker awards

Pioneers in anaesthesiology, DNA packaging in chromosomes and RNA biology are this year's recipients of the Lasker awards. The annual biomedical research prize, announced on 11 September, is often considered a precursor to the Nobel prize. John Glen,

a scientist formerly at the drug company AstraZeneca in London, won the clinical award for his discovery of the anaesthetic propofol. Joan Argetsinger Steitz of Yale University in New Haven, Connecticut, won the special achievement award for her work on RNA and for promoting women in science. And the awardees for basic science were C. David Allis of the Rockefeller University in New York City and Michael Grunstein of the University of California, Los Angeles, who discovered how chemical modifications to certain proteins on chromosomes

can turn genes on and off. Each award comes with a US\$250,000 honorarium.

Chinese accolades

Yuan Longping — the 'father of hybrid rice' — and two other rice researchers have won the Future Science Prize for the life sciences. Yuan shares the award with Zhang Qifa and Li Jiayang from the Chinese Academy of Sciences. The prizes — often called China's Nobels — honour discoveries made in the nation, and each of the three awards is worth US\$1 million. Ma Dawei at the Shanghai Institute of Organic Chemistry, Feng Xiaoming at



ALAMY

France admits role in lecturer's murder

French President Emmanuel Macron has officially recognized that the French army tortured and killed mathematician Maurice Audin during the Algerian War of Independence. Macron presented a statement to Audin's 87-year-old widow, Josette, at the family home near Paris on 13 September. Audin was a French university lecturer in Algiers and a member of the Algerian

Communist Party. He disappeared in June 1957, and the authorities claimed that he had escaped after being arrested for harbouring communist-party members. Macron's statement also acknowledged — for the first time — that the French army systematically used torture during the war. It added that all records for people who disappeared during the conflict would be made open.

KIM SHIFLETT/NASA
Sichuan University and Zhou Qilin at Nankai University shared the physics prize for discovering catalysts that can synthesize drug molecules. Lin Burn, formerly at the Taiwan Semiconductor Manufacturing Company in Hsinchu City, took home the mathematics award for improving a technique to manufacture integrated circuits.

SPACE

Satellite launch

NASA's Ice, Cloud, and land Elevation Satellite-2 (ICESat-2) launched from the Vandenberg Air Force Base in California on 15 September (pictured), kicking off a three-year mission focused on tracking ice thickness. The satellite will also measure cloud heights and forest growth. The craft's Advanced Topographic Laser Altimeter System (ATLAS) can track changes in ice thickness to within half a centimetre per year. The data it collects will provide seasonal and annual information that can better inform predictive models of melting ice and rising sea levels. Researchers conceived the mission in 2008, but it took about five years to develop ATLAS. Technical difficulties and budget



considerations then delayed the ICESat-2 launch until this year.

Space agency

Luxembourg launched a national space agency on 12 September. Nestled in the country's economic ministry, it is aimed at advancing space initiatives in business rather than building and launching spacecraft. Luxembourg has long been involved in the space industry, and is the base for commercial satellite operators such as Intelsat and SES. In 2016, the country established a space-mining initiative to explore who would own minerals extracted in space. The Luxembourg Space Agency will involve business, academia, non-profit organizations and other partners to promote economic

development in space. It aims to create a €100-million fund (US\$117 million) to help finance new space businesses.

POLICY

Methane proposal

On 11 September, the US Environmental Protection Agency (EPA) announced a proposed rule that would relax regulation of methane emissions at oil and gas facilities. The plan would roll back key provisions of the methane regulations put in place under former president Barack Obama in 2016. The plan includes a reduction in the frequency of emissions monitoring at oil and natural-gas wells, and at facilities that compress gas for transport through pipelines. It also extends the time that companies have to fully repair leaks of this potent greenhouse gas from 30 days to 60 days. The EPA is soliciting comments for 60 days, and plans to hold a public hearing on the proposed rule in Denver, Colorado.

Harassment policy

The US National Institutes of Health (NIH) is rethinking how it handles sexual harassment. The agency will soon introduce a centralized system for reporting

harassment by NIH scientists, director Francis Collins said on 17 September. The agency has launched an anti-sexual-harassment website. It will also update its harassment policy, launch training and education campaigns to prevent harassment, and survey its staff and contractors about workplace climate and harassment issues. The moves come several months after the National Science Foundation announced that it would update its harassment policies.

EVENTS

Cochrane board

The 13-member board of the prestigious medical-evidence group the Cochrane Collaboration has seen a slew of resignations. On 14 September, board member Peter Gøtzsche said he had been expelled from the collaboration with no clear justification. Cochrane told *Nature* that it had received "numerous complaints" about Gøtzsche, after he co-authored a critique of the Cochrane's review on the human papilloma virus vaccine. Four elected board members quit in protest, requiring two appointed members to step down to maintain a mandated balance. The ructions came as the group had its annual meeting in Edinburgh, UK.

SOURCE: OECD

TREND WATCH

Ten years ago, the world's economy came crashing down — a crisis encapsulated by the demise of the US investment bank Lehman Brothers, which filed for the largest bankruptcy in history on 15 September 2008. A decade on, *Nature* looks at how funding for research — which relies on both government and private investment — has fared around the world.

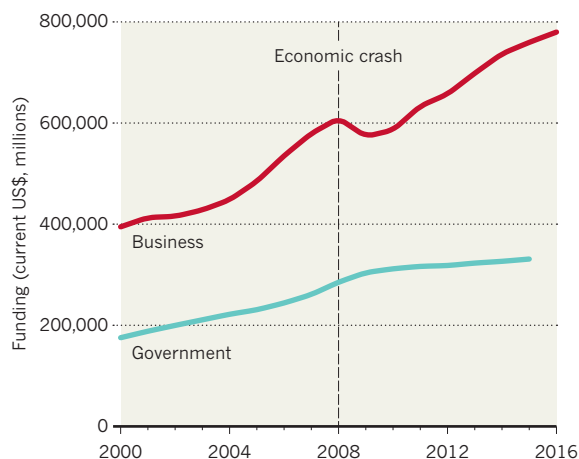
Data from the Organisation for Economic Co-operation and Development show that there has been a steady growth overall in research and development (R&D) funding since the initial aftermath of the crash. Among

the organization's 36 countries — which include the United States, Japan and many European nations — R&D funding from businesses dipped briefly after the crash, but then recovered, and has climbed since. Government funding generally rose in the two years after the crisis, possibly reflecting countries' attempts to stimulate their economies.

But disparities exist between nations. In European countries hit hardest by the recession, such as Greece and Spain, R&D funding from the government plummeted and has not fully recovered. See go.nature.com/2xg7afi for more.

TEN YEARS AFTER THE CRASH

Although research and development funding by the business sector dipped after the 2008 economic crash, it soon recovered in Organisation for Economic Co-operation and Development countries.



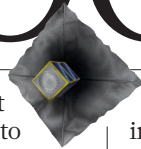
NEWS IN FOCUS

PHYSICS Machine learning helps unlock 'dark matter' of bizarre superconductors **p.294**

ECOLOGY The hidden lives of deep-sea creatures can now be caught on camera **p.296**

SPACE Satellite tests net and harpoon as ways to clean up space junk **p.297**

NEUROSCIENCE Voltage imaging provides way to capture nerve behaviour **p.300**



DAVID GRAY/REUTERS



Large parts of Australia are enduring a crippling drought.

POLICY

Australia abandons plan to cut carbon emissions

Scientists say the move amounts to walking away from the Paris climate agreement.

BY ADAM MORTON

Australia's new prime minister has abandoned the country's policy to cut greenhouse-gas emissions. Climate scientists say the move means the government has effectively dropped its commitment to the 2015 Paris climate agreement.

"They've walked away from Paris without saying it, hoping no one would notice," says Lesley Hughes, a climate-change scientist at Macquarie University in Sydney.

Australia now becomes the second advanced

economy, after the United States, to drop emissions-reduction policies since the 2015 Paris climate conference. US President Donald Trump signed an executive order to start removing climate regulations in March 2017 and pulled out of the Paris agreement in June that year.

Australia's effective abandonment of Paris can be traced back to late August, when the ruling conservative Liberal Party abruptly replaced former leader Malcolm Turnbull with Prime Minister Scott Morrison. The leadership change came after some party members objected to a policy, the National Energy

Guarantee (NEG), that would have required electricity companies to meet emissions targets. Morrison subsequently said that he was abandoning the NEG, and would instead focus on reducing the cost of energy for the public.

The abandonment comes as large parts of country feel the effects of global warming — a crippling drought grips the eastern states and dozens of bushfires have erupted unseasonably early in those regions.

Some government members have even suggested that the country should officially withdraw from the Paris agreement. ►

► Morrison has rejected this idea. He says Australia is on track to meet the target it announced before the Paris conference: to cut emissions by 26–28% below 2005 levels by 2030.

But there is little evidence to suggest the government will be able to meet this target without new policies. In August, government advisers said it was unlikely that the electricity sector, responsible for one-third of Australia's emissions, would reduce its emissions by 26% unless a policy was introduced to drive cleaner energy generation over the next decade.

ON THE RISE

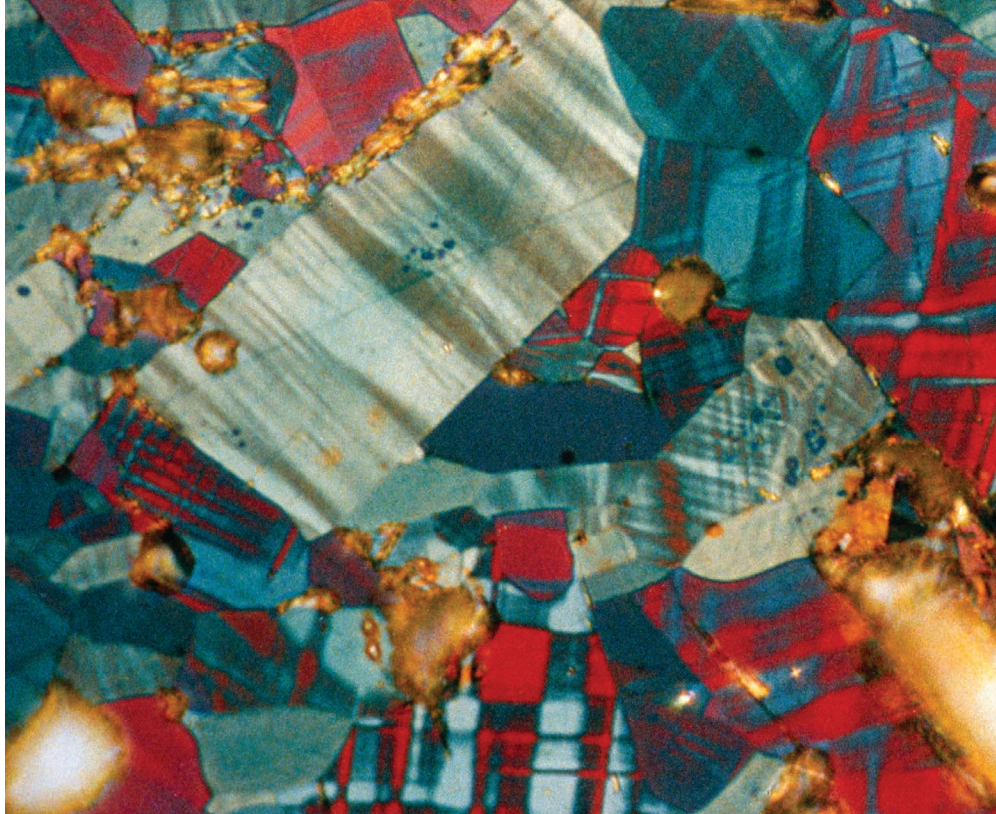
National emissions have risen each year since 2014, when the government repealed laws requiring big industrial emitters to pay for their emissions. There are also no significant policies to reduce the other major sources of pollution, such as transport, agriculture, heavy industry and mining, which together generate nearly two-thirds of Australia's carbon emissions.

Although the NEG was a modest policy, proposed after several more-effective schemes failed to win political support, it had the potential to win the backing of the centre-left opposition Labor Party, says John Church, a specialist in sea-level rise at the Climate Change Research Centre (CCRC) at the University of New South Wales in Sydney. That would have enabled the policy to pass through parliament and into law. The policy also had the support of the business community, which has been calling for climate and energy strategies that encourage investment in new and cleaner power plants, he says. "Walking away from it was a disaster"

Sarah Perkins-Kirkpatrick, an authority on heatwaves, who is also at the CCRC, says government motivation to do something about climate change seems to have disappeared altogether. When she briefed senior officials on the latest climate-change science in August, she left the meeting feeling optimistic that more policies were coming. "People were trying to get things done, but now that's not the case at all," she says. "I'm extremely frustrated."

The decision to drop the policy also goes against the public's support for action on climate change, says Hughes. A poll of 1,756 people, published on 12 September by research and advocacy organization the Australia Institute, found that 73% of respondents were concerned about climate change and 68% wanted domestic climate targets in line with the country's Paris commitment.

But Australia's lack of climate policy could be short-lived. A national election is due by May 2019, and recent polls suggest that the Labor Party, led by former union boss Bill Shorten, is favoured to win. Labor says it would set a new emissions target of a 45% cut by 2030, although it has not revealed how it would reach this goal. In the meantime, some states have mandated ambitious renewable-energy targets, and business leaders say investment in clean energy is increasing because it is now the cheapest option. ■



PHYSICS

AI spots pattern in superconductor data

Machine learning might one day boost efforts to make sense of other baffling quantum systems.

BY ELIZABETH GIBNEY

Machine-learning algorithms are helping to unravel the quantum behaviour of a type of superconductor that has perplexed physicists for decades.

Researchers used artificial intelligence (AI) to spot hidden order in images of a bizarre state in high-temperature superconductors.

The result, published in a preprint¹ on the arXiv server last month, supports one theory in a decades-long attempt to understand these materials.

The study also represents the first time that machine learning has been successfully used to make sense of experimental data on quantum matter, said Eun-Ah Kim at Cornell University in Ithaca, New York, who presented the work at a meeting on Materials and Mechanisms of Superconductivity and High Temperature Superconductivity in Beijing in August.

In the long term, machine learning might boost efforts to spot simple patterns in other noisy and chaotic experimental systems, such as quantum spin liquids, which could form the basis of a future exotic type of quantum computer.

The latest study focused on superconductors, which conduct electricity without any electrical resistance, but typically do so at less than 4 degrees above absolute zero, around –269 °C. Kim's team examined an even more rarefied group, called cuprates, which are made of sandwiches of copper oxide and become superconducting at temperatures up to –140 °C. Understanding the reason that cuprates can superconduct could be the key to engineering materials that do so closer to room temperature.

MYSTERY STATE

But things get particularly baffling when cuprates enter a state called the pseudogap, which occurs when the materials are close to superconducting. Complex interactions between electrons and atoms make the pseudogap difficult to describe theoretically and its chaotic nature challenging to observe. Some physicists call this state the cuprates' 'dark matter', yet explaining the pseudogap may be key to understanding superconductivity.

Physicists have observed some promising signs of order in the pseudogap — visible as ripples of changing electron density — but

SHEFFIELD UNIV./P. WARD & T. BUTTON/SPL



Micrograph of yttrium barium copper oxide, a high-temperature superconductor.

studied images of the pseudogap. Patterns in the images, taken with a scanning tunnelling microscope, often seem disordered to the human eye because of the material's naturally chaotic and fluctuating nature, and noise in the measurements. The advantage of machine learning in this situation is that algorithms can learn to recognize patterns that are invisible to people.

PATTERN RECOGNITION

To train the algorithms, the team fed neural networks examples of rippled patterns that corresponded to different theoretical predictions. Having learnt to recognize these examples, each algorithm applied this learning to real data from cuprates in the pseudogap. Over 81 iterations, the algorithms repeatedly identified one modulating pattern that corresponded to the particle-like description of electrons, which dates back to the 1990s.

The team's paper shows that the particle-like description is more appropriate in this case than is the conventional wave-like description, says André-Marie Tremblay, a physicist at the University of Sherbrooke in Canada, who was at Kim's talk in Beijing. Working out the nature of the patterns is crucial to interpreting what causes them, says Milan Allan, a physicist at Leiden University in the Netherlands.

The technique could eventually help physicists to understand high-temperature superconductivity, says Allan, although he cautions that the paper is far from definitive and that debate about what the pseudogap is will continue.

The work is an impressive, original application of machine-learning algorithms to this type of experimental data, says Tremblay. But the algorithm can only distinguish between the various hypotheses it is given, he says, rather than find entirely new patterns.

During her talk, Kim said that work is under way to apply the technique to rapidly make sense of data from the X-ray diffraction of quantum materials — a technique that uses the scattering of electromagnetic waves to reveal a material's 3D physical structure, but which creates patterns so rich that they can take months to unravel by conventional means. In this case, the AI must draw out similarities and classifications itself, rather than be given pre-labelled examples, by grouping features that it sees as similar. "This journey of using AI, or machine learning, for various aspects of our quest to understand quantum emergence has just begun," said Kim. ■

1. Zhang, Y. et al. Preprint at <https://arxiv.org/abs/1808.00479> (2018).
2. Kivelson, S. A., Fradkin, E. & Emery, V. J. *Nature* **393**, 550–553 (1998).
3. Zaanen, J. *Science* **286**, 251–252 (1999).

they disagree on how to explain these patterns. One approach views electrons as strongly interacting particles^{2,3}, whereas the other treats them as wave-like and only weakly interacting.

To glean more information about these patterns, Kim's team designed neural networks — AI inspired by structures in the brain — that

GENDER BIAS

Peer review fails equity test

Analysis of submissions to eLife reveals a gender gap in whom journals invite to do reviews.

BY DALMEET SINGH CHAWLA

Women are inadequately represented as peer reviewers, journal editors and last authors of studies, according to an analysis of manuscript submissions to an influential biomedical journal.

The study looked at all submissions made to the open-access title *eLife* from its launch in 2012 to 2017 — nearly 24,000. It found that women worldwide, and researchers outside North America and Europe, were less likely to be peer reviewers, editors and last authors. The paper — which hasn't itself yet been peer-reviewed — was posted on the preprint server bioRxiv on 29 August (D. Murray et al. Preprint at BioRxiv <https://doi.org/10.1101/400515>; 2018).

About 7,000 of the submitted studies went through the full submission process (at *eLife*, authors make a 'pre-submission query' before being invited by the journal to send a full paper — a relatively uncommon practice among

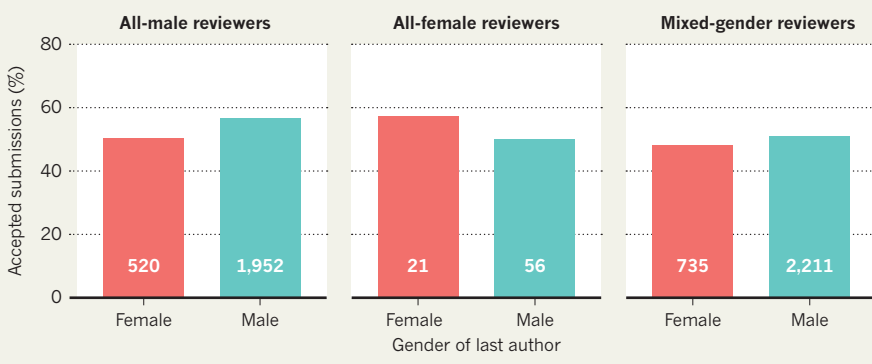
journals). In all, the analysis covered the activity of about 7,000 referees, 890 reviewing editors and 57 senior editors.

The researchers found that women make

up only 21% of peer reviewers, and around one in four reviewing editors. Most reviewing editors and peer reviewers were based in the United States — 62% and 56%, respectively ▶

PEER-REVIEW PATTERNS

An analysis of thousands of submissions to the journal *eLife* — in which peer-review panels openly discuss submitted works — found that all-female reviewer groups accepted more manuscripts with female last authors than did all-male panels.



► — followed by the United Kingdom and Germany in second and third place. Less than 2% of reviewers were in developing nations.

Of the full submissions, the study found that 1,549 (22%) had a female last author — a position that indicates seniority — and 5,127 had a male last author. About 53% of manuscripts with male last authors were accepted, compared with around 50% of those with female last authors.

Fifty-seven per cent of fully submitted papers with a male last author were accepted when the review panel was all male (see ‘Peer-review patterns’), whereas mixed-gender teams accepted 51% of male-last-author papers. And submissions that had been edited or reviewed by someone in the same country as the corresponding author were more likely to be accepted than those with a country mismatch.

The trends are likely to be a result of implicit biases, says study co-author Cassidy Sugimoto, an information scientist at Indiana University Bloomington. The study did not seek to reveal how the disparities arose, say the authors. But because the gender make-up of senior authors and gatekeepers closely matches disparities found broadly in science, there is no evidence that *eLife* is making such disparities worse.

The research was prompted by *eLife*, which approached Sugimoto and her colleagues with the data; two study authors are *eLife* employees. The journal’s reviewing process is unorthodox

in that referees know each other’s identities, which allows them to discuss any differences of opinion on manuscripts.

BODY OF EVIDENCE

The study is robust, says Jevin West, an information scientist at the University of Washington in Seattle. And it is concerning that women and authors in developing countries seem to be marginalized in peer review, he says. “It’s very important that we have diverse voices represented and that those voices are treated equitably.”

The results echo previous findings about peer review. This month, a global survey by Publons — a site that allows academics to record their peer-review activity — found that researchers in developing countries are under-represented as reviewers, yet are more likely than scientists in richer countries to accept review requests, and complete reviews faster.

And last year, an analysis of American Geophysical Union (AGU) journals found that women are invited to review less often than expected, but that the editors’ gender has no influence on acceptance rates (J. Lerback and B. Hanson *Nature* **541**, 455–457; 2017).

“It’s important that we have diverse voices represented and that those voices are treated equitably.”

Sugimoto says that journal policies should aim to ensure diversity on review panels, for example, by inviting a greater proportion of women and researchers in developing nations to do reviews. “This is one of the simplest policy changes we can make,” she says, “without high risks, and potentially high benefits.”

Andrew Collings, *eLife*’s executive editor and a study co-author who is based in Cambridge, UK, says that the team is communicating its results to the editorial board, so that editors can consider the findings as they assess submissions and select reviewers. “We are particularly keen to see editors using diverse groups of reviewers whenever possible.”

To weed out the effect of implicit biases on acceptance rates, it is tempting to see blinding as a solution, West says. But, he adds, double-blind peer review — in which neither authors nor reviewers know each other’s identities — often works poorly, because some fields are so small that reviewers can guess who wrote a paper.

Sugimoto says that more data are needed to determine the effectiveness of techniques such as blinding or open peer review, in which reviews are published and authors and reviewers might know each other’s identities.

She hopes that more journals and publishers will release data on peer review for analysis. “Then, we can inform it with evidence rather than with anecdote.” ■

ECOLOGY

Hidden lives of deep-sea animals

Cameras record behaviours long cloaked in darkness.

BY AMY MAXMEN

Advan­ces in video cameras and low-light sensors are revealing animal behaviours in the deep sea that researchers have never recorded before.

The behaviours include a worm-like predator shooting off rings of blue light, and an animal anchored to the sea floor sending flashes of light dancing along its body, creating the illusion of a tiny creature swimming upwards.

Steven Haddock, a marine biologist at the Monterey Bay Aquarium Research Institute (MBARI) in California, showcased videos of these phenomena and more for the first time on 13 September at the Deep Sea Biology Symposium in Monterey. He is one of a handful of researchers around the world who are

using extremely high-resolution cameras and ultra-sensitive sensors to capture unprecedented footage of marine organisms in the wild.

“We can see natural behaviour in a way that we’ve never been able to before,” says Haddock.

COMING INTO FOCUS

Until recently, researchers needed to use bright lights to capture footage of animals living in the deep dark ocean. The lights scared many creatures away, and when scientists tried filming under low-light conditions, poor camera resolution made it difficult to pick out fine details such as a small ring of light.

In 2016, Haddock’s team attached a 4K camera, which has four times as many

pixels per image as a high-definition (HD) camera, to one of MBARI’s remotely operated vehicles (ROVs). On one of Haddock’s first voyages with the camera, he recorded a 2.5-centimetre-long animal called an arrow worm emitting a trail of doughnut-shaped rings of blue light. Haddock speculates that the creature uses the display to distract predators as it escapes. “Our HD camera wouldn’t have captured this at all,” he says.

In mid-August, another research team deployed an 8K camera in the deep sea for the first time to explore hydrothermal vents in the Okinawa Trough near Japan. The 8K camera’s resolution nearly matches that of the human eye, and it enabled Dhugal Lindsay, a marine biologist at the Japan Agency for Marine-Earth Science and Technology in Yokosuka, to film near-microscopic plankton in enough detail to identify their species.

SEEING IN THE DARK

Other marine biologists are fine-tuning the latest low-light camera sensors that also reduce noise from scattered, indirect light. This allows researchers to use a lot less illumination to record ocean life, decreasing the chances of their ROVs scaring off animals.

The sensors also allow scientists to pick up phenomena such as bioluminescence — the production of light by an organism — and to



Dim red light illuminates a bioluminescent display by an *Atolla* jellyfish.

HADDOCK/MBARI

identify the animals giving off the light show.

“I can’t tell you how many times I’ve seen bioluminescence in the dark and said, ‘hey, that was cool, but I have no idea what it is,’” says Brennan Phillips, an oceanographer at the University of Rhode Island in Narragansett.

A few years ago, Phillips recorded an as-yet unidentified species of *Tomopteris*, a marine worm that looks like a centipede, using cameras fitted with advanced low-light sensors. He was able to capture footage of light glowing in the animal’s central nervous

system and then radiating into each of its legs.

And on a trip off the coast of Mexico in May, Phillips and other researchers used another new, specialized sensor to record an elusive 68-centimetre-long jellyfish called *Deepstaria enigmatica* (D. F. Gruber *et al.* *Am. Mus. Novit.* No. 3900; 2018). This jellyfish lacks tentacles, and researchers had long wondered how it captured its prey. The detailed footage showed how the invertebrate moved, which enabled scientists to deduce that the animal ‘bags’ its meal using the thin, membrane-like sac of its body.

Roughly three-quarters of marine organisms, excluding microscopic species and those that live on the sea floor, produce light. But researchers are only beginning to learn how the creatures use this ability to communicate, to attract mates or prey, or to defend themselves, says Haddock.

The footage that he and others are collecting shows animals acting in ways scientists have never before recorded, prompting more questions than answers. “We are going deeper than ‘gee-whiz,’” Haddock says. ■

TECHNOLOGY

Harpoon-throwing satellite takes aim at space junk

Tests of experimental craft include flinging a net and shooting a spear at targets in space.

BY ALEXANDRA WITZE

In a move Spiderman might envy, one satellite flung a net at another craft in low Earth orbit on 16 September. A few months from now, the satellite will ape the spear-wielding Aquaman and fire a harpoon into space.

The manoeuvres will test ideas meant to address the growing problem of space junk. If they work, future missions might use similar nets or harpoons to ensnare dangerous space debris and drag it to a fiery end in Earth’s atmosphere.

“This is proof of concept of a new technology,” says Guglielmo Aglietti, director of the Surrey Space Centre at the University of Surrey in Guildford, UK, and principal investigator for the project, known as RemoveDEBRIS. “The idea is to be really useful and clean up satellite space.”

The US military tracks approximately 20,000 objects in orbit that measure at least 5–10 centimetres across. That’s big enough to cause serious damage if two objects collide, and the threat is growing as more junk builds

up in space. In 2009, a US communications satellite accidentally smashed into a Russian one — creating thousands of shards that now hurtle through low Earth orbit, raising the threat of future collisions.

Now researchers are dreaming up ways to clean up some of this orbital junk. Last year, the Japan Aerospace Exploration Agency tried to unfurl an electrodynamic tether and hook it on to a piece of space debris; the mission failed when the tether did not release as expected. A team spun off from the Swiss Federal Institute of Technology in Lausanne (EPFL) is raising money to build a satellite that would throw a conical net around a defunct craft and steer it to its doom. And the European Space Agency (ESA) is working on ideas for a more complex spacecraft that could dispose of space junk or perhaps even refuel a satellite in

orbit, extending its life, says Luisa Innocenti, head of ESA’s Clean Space initiative in Paris.

TAKING OUT THE TRASH

The €15-million (US\$17-million) RemoveDEBRIS mission is meant to test cheap ways to drag junk out of orbit. “There will always be a tension between letting debris stay as it is or going to clean up some of it,” says Aglietti. But if a space agency could remove particularly big and dangerous pieces of debris — such as ESA’s defunct, bus-sized Envisat Earth-observing satellite — it might be worth the effort.

RemoveDEBRIS will test four technologies over a carefully choreographed few months. The spacecraft launched to the International Space Station in April and deployed into space in June. The first test, the net experiment, took place on 16 September (see ‘Ready, aim... fire’).

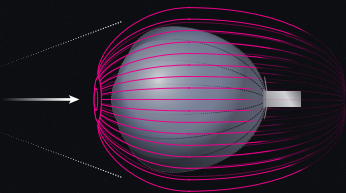
The craft ejected a CubeSat, a satellite about the size of a loaf of bread, which inflated a balloon to a diameter of roughly 1 metre — big enough to be worth grappling with. RemoveDEBRIS then hurled its net around the ►

“There will always be a tension between letting debris stay as it is or going to clean up some of it.”

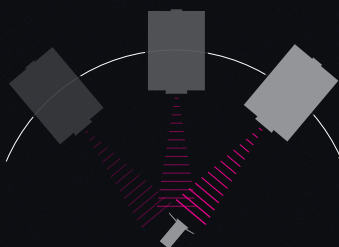
READY, AIM... FIRE

The €15-million (US\$17-million) RemoveDEBRIS mission is testing ways to clean up space junk, beginning this month.

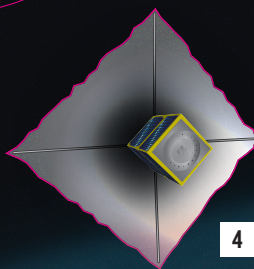
- 1** On 16 September, the satellite threw and cinched a net around a 1-metre-wide balloon target.



- 2** In a few weeks, it will deploy a tiny satellite and use lasers to analyse its movement.



- 3** In early 2019, it will extend an arm and shoot a harpoon at a target 1.5 metres away.



- 4** Finally, it will unfold a sail to drag it into Earth's atmosphere, where it will burn up.

► balloon, using weights to cinch the net closed like a purse.

"It went well," says Aglietti. "We are very happy."

The second experiment is planned for late October. RemoveDEBRIS will eject a second CubeSat, then scan it using lasers to test techniques for studying and

navigating near space junk.

The third test, throwing the harpoon, could come in early February. The RemoveDEBRIS satellite will extend an arm 1.5 metres into space, flip up a target plate and shoot the spear at it.

Finally, in March, the satellite will inflate a 1-metre-long mast and unfurl a sail. The sail is meant to function as a drag, steering the satellite to lower altitudes so it will eventually burn up in the atmosphere.

An industry consortium built the project, with subsidiaries of the aerospace company Airbus providing the net and the harpoon. The team has tested each experiment on the ground, but things could go awry in the notoriously difficult environment of space.

"We are very much prepared for some of the things to go a bit differently than planned," says Aglietti. But if it works, the space net and harpoon could become common weapons for dealing with space junk. ■

CORRECTION

The News feature 'The information factories' (*Nature* **561**, 163–166; 2018) erroneously affiliated Eric Masanet with Northeastern University. In fact, he is at Northwestern University in Evanston, Illinois.



THE BRAIN'S RAUCOUS SYMPHONY

New precision proteins can make high-fidelity recordings of neuronal activity. The results could reveal how circuits generate thoughts and emotions.

BY GIORGIA GUGLIELMI

Biophysicist Adam Cohen was strolling around San Francisco, California, in 2010, when a telephone call caught him by surprise. “We have a signal,” said the caller. Nearly 5,000 kilometres away, in Cambridge, Massachusetts, his collaborators had struck gold. After months of failed experiments, the researchers had found a fluorescent protein that allowed them to watch signals as they passed between neurons.

But there was something weird going on.

When Cohen got back to his lab at Harvard University, he learned that all the recordings of the experiment showed a strange progression. At first, neurons decorated with the protein flashed nicely as electric impulses whizzed through them. But then the cells turned into bright blobs. “Halfway through each recording, the signal would go all wild,” Cohen says.

So he decided to join his team during an experiment. “When they started the recording, they would sit there holding their breath,”

ILLUSTRATION BY JOANNA GEBAL

Cohen says. But as soon as they realized it was working, they would celebrate, “dancing and running around the room”.

In their exuberance, they were letting the light from a desk lamp shine right onto the microscope. “We were actually recording our excitement,” says Daniel Hochbaum, then a graduate student in Cohen’s group. They toned down their celebrations, and a year later, the team published its study¹ — one of the first to show that a fluorescent protein engineered into specific mammalian neurons could be used to track individual electric impulses in real time.

Neuroscientists have tried for decades to observe the swift electrical signals that are a major component of the brain’s language. Although electrodes, the workhorse for measuring voltage, can reliably record the activity of individual neurons, they struggle to capture the signals of many, particularly for prolonged periods. But in the past two decades, scientists have found a way to embed fluorescent, voltage-indicating proteins right into the cell membranes of neurons. With the right kind of microscope, they can then see cells lighting up as they talk to each other — be it in a whisper or a shout. Voltage imaging can also record electrical chatter between many neurons at once, and then average those signals across large chunks of brain tissue. This helps researchers to study the brain’s electrical activity across different spatial scales, by listening not only to the voices of individual cells but also to “the roar of the crowd”, Cohen says.

In the past 5 years, scientists have published about 1,000 papers on the topic, and major funding schemes such as the US National Institutes of Health’s BRAIN initiative have sped up the development of new types of genetically engineered voltage indicators. In the hope of finding better variants, some groups have come up with strategies to screen millions of proteins for desired characteristics such as brightness. One such approach has identified an indicator that’s twice as bright as similar sensors developed just four years earlier².

As these proteins improve, and advances in microscopy make it easier to see them, scientists hope to illuminate neuroscience’s biggest puzzle: how the brain’s cells work together to transform a system of electrical pulses into thoughts, actions and emotions. Researchers are still struggling to catch the full range of activity and to devise ways to see nerves firing fast and deep within brain tissue. But if advances can solve these technical challenges, “it would be revolutionary”, says Rafael Yuste, who studies the function of neural circuits at Columbia University in New York City.

HIGH-SPEED PROCESS

The average human brain contains about 120 billion neurons, which constantly receive and send information through branch-like appendages called dendrites. Chemical or electrical signals that reach the dendrites produce small voltage changes across the cell’s membrane, which are routed to the cell body. When the sum of the voltage changes reaches a point of no return, called a threshold, the neuron fires a large electrical spike — an action potential. This jolt whizzes at speeds of up to 150 metres per second along a neuronal branch, known as an axon, to another set of branching appendages. Here, chemical or electrical signals pass the information on to the next set of dendrites.

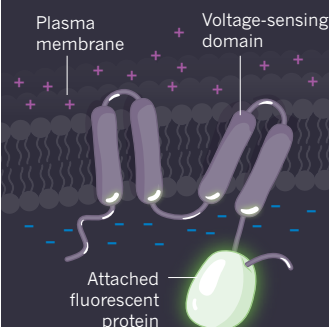
Neuronal signals converge, diverge and synchronize to produce a symphony of thoughts, emotions, actions and reactions, from the flush of a face to a baby’s hiccup. But scientists’ listening tools are extremely limited. First developed in the 1940s, miniature electrodes as thin as a hair can be inserted into the brain, up against or inside neurons, where they measure membrane voltage with precision and speed. But this approach can be used to monitor just one or a handful of neurons at once — and only for a limited amount of time, because the electrodes eventually damage the cell. It’s like trying to get the gist of an orchestral arrangement by following one player for a few seconds.

Bundles of micro-electrodes can record the electrical activity of up to 200 cells at once, but because these electrodes are placed near to neurons rather than inside them, they can detect only the action potentials, the sharpest spikes in electrical activity. They are deaf to softer notes — the small electrical changes that do not push the neuron

FLAVOURS OF FLUORESCENCE

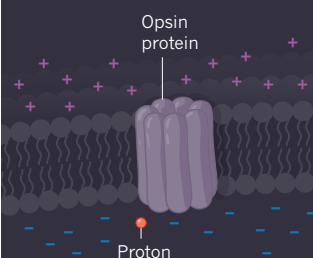
Scientists have built different types of genetically encoded voltage indicator (GEVI). One major category (top) uses a membrane-bound portion of a voltage-sensing protein, such as a sodium channel, fused to one or more fluorescent proteins. Another category uses an opsin protein, such as a microbial rhodopsin, a membrane channel that directly changes its fluorescent properties in response to an electric field.

Voltage-sensing fusion protein



A voltage change across the membrane causes the GEVI to change shape, decreasing the fluorescence of the attached protein.

Opsin-based voltage indicator



A voltage change across the membrane can help to add a proton to retinal, the light-sensitive portion of the opsin, which alters its fluorescence.

all the way to an action potential. These sub-threshold voltage changes are key to brain function, because they gradually add up to determine whether or not a neuron will fire.

In the hope of measuring quieter brain activity in larger populations of cells, scientists in the 1960s began toying with the idea of a sensor or probe that fluoresces in response to an electrical signal. The most popular probes, called calcium indicators, light up when they bind to calcium, which flows into the neuron as a result of a spike in electrical activity. But the technique, known as calcium imaging, provides only a proxy; it doesn’t directly record membrane voltage. And although it will show the signal of big events such as action potentials, it misses things that are crucial for brain function, such as subtle swings in membrane voltage or the electric signals that inhibit action potentials. Imagine being able to hear only a burst of applause after a symphony concert: it’s clear that the orchestra has performed, but what it was playing is anyone’s guess.

In the 1970s, scientists started to develop dye sensors that detect changes in membrane voltage directly. The first versions of these dyes had to be painted onto the brain indiscriminately, so they labelled all cell types, including non-neuronal cells, making it difficult to parse the activity of specific neurons.

Then, in the 1990s, researchers started testing indicators that could be genetically engineered to show up only in neurons of interest. The first³ genetically encoded voltage indicator (GEVI) was developed in 1997; since then, scientists have churned out more than two dozen sensors⁴. Some of these are made by combining a voltage-sensitive protein with fluorescent molecules (see ‘Flavours of fluorescence’). When these proteins detect a change in voltage, they change their 3D structure and alter the fluorescence of the molecule they’re coupled to. Other voltage indicators are mutated versions of microbial rhodopsins, fluorescent molecules that cause a change in voltage across the plasma membrane

in response to light. These proteins can also work in reverse, changing their response to light — and thus their fluorescence — in response to a change in membrane voltage.

ALL IN THE DETAIL

So far, GEVIs have proved successful in tracking individual action potentials in both cultured neurons, grown in a dish, and in the intact brains of a wide range of animals, from insects⁵ to mice⁶. One of the biggest promises of the technique is its potential to record not only the big events but also the small, sub-threshold changes in membrane voltage that reflect the messages that a neuron receives from neighbouring cells, Cohen says. “Voltage imaging lets you see the inputs to the neurons *in vivo*, which we had no way to look at previously,” he says.

In the past year, Cohen and his colleagues developed new GEVIs and improved microscopy techniques to record such sub-threshold voltage changes from many neurons at once, including in the mouse brain^{7,8}. The team was also able to record the electrical activity of the same neurons up to a week later. The ability to know exactly which neurons are being recorded and to keep track of them over time allows researchers to look at the wiring between those neurons, says Ed Boyden, a neuroscientist at the Massachusetts Institute of Technology in Cambridge. By doing so, “you can link the structure of the brain with its function,” he says. “That’s one of the core questions in all neuroscience.”

Another advantage of GEVIs is that, unlike electrodes, which record mainly signals from the cell body, they can record electrical signals from any part of a nerve cell, right down to the tips of dendrites. That’s like being able to listen specifically to the notes played by a pianist’s left hand. “This is something that I’ve been dreaming for a long time — and I’m not alone,” says Katalin Toth, a neurobiologist at Laval University in Quebec City, Canada. Many neuroscientists are striving to follow voltage across entire neurons to see how it changes in different regions of the cell, she says.

Wei Wei, a neurobiologist at the University of Chicago, Illinois, is using GEVIs to work out how different electrical inputs are integrated in the neurons of the mouse retina. Wei is interested in a class of neuron that responds more strongly to a visual stimulus when it is moving in a particular direction. By looking at how membrane voltage changes in different parts of these neurons, she hopes to understand how the cells sum up incoming signals to detect the direction of the movement.

Neurophysiologist Vincent Villette at the Ecole Normale Supérieure in Paris plans to use voltage sensors to study how regular fluctuations of sub-threshold electric signals determine how neurons in the mouse cerebellum coordinate muscle activity. “There’s a lot to be understood on how cells act together,” Villette says.

Getting a visual read-out of membrane voltage also allows scientists to see electrical signals that inhibit neuronal firing rather than trigger it. Because inhibitory signals are impossible to record with approaches such as calcium imaging, it’s unclear how exactly they shape brain activity, says Rosa Cossart, a neurobiologist at the Mediterranean Institute of Neurobiology in Marseilles, France.

Cossart has been using electrodes and calcium imaging for years, but she’s now eager to try GEVIs. She hopes these sensors will allow her to measure voltage at high speed across multiple neurons — at least 50 — at the same time in a living mouse. This would help to understand how groups of neurons integrate electrical signals — both excitatory and inhibitory — to support activities that are crucial for brain development and function, she says.

DEEP CHALLENGES

Despite the high expectations, getting GEVIs to work in the laboratory can be a hassle. Take Helen Yang: as a graduate student at California’s Stanford University, she decided to try GEVIs as a way to study neurons in the fruit fly’s visual system. But peering through the microscope during her first experiment, Yang saw no change in the cells’ fluorescence, not even when

she flashed a bright light in the flies’ eyes. It wasn’t until she analysed the data that she realized that the visual stimuli were producing a signal, it was just a tiny one. “I was pretty excited, but my lab-mates were less so,” she says. “The responses were pretty small and noisy.”

Yang started to play with the microscope settings, increasing the laser power and speeding up the imaging. “I basically made it go as fast as our microscope could,” she says. That’s because the indicator’s response to an electrical signal was so fast that the change in fluorescence was detectable just for a fraction of a second. “If you’re only capturing one frame during the time that the cell is responding, the response doesn’t look big at all,” Yang says.

Yang eventually managed to use GEVIs to investigate how the flies’ neurons process visual cues⁵, but the sorts of challenges she faced have so far prevented voltage imaging from becoming a mainstream technique. It requires advanced, often custom-built microscope platforms, Cohen says. “You can’t just do this on your grandmother’s fluorescent microscope.”

In the past five years, financial support from the BRAIN initiative has boosted advances in the field, including the development of better GEVIs, says Michael Lin, a protein engineer at Stanford.

In parallel with the development of new sensors, scientists are working on techniques to image with precision the fast electrical signals travelling through the brain. One challenge is that most of the available techniques work well only with cells in a dish or on the surface of the brain. But

the mammalian brain isn’t transparent: in fact, it looks like tofu, says Na Ji, a physicist at University of California, Berkeley.

To peer deeper, researchers have to turn to more-invasive methods, such as removing some of the overlying tissue or sticking tiny optical devices called micro-endoscopes directly into the brain. An alternative, non-invasive way to look into opaque tissues — up to 1 millimetre deep — is two-photon microscopy. This technique uses longer-wavelength, lower-energy light, which can penetrate deeper into tissues. Because two-photon

microscopes illuminate and record from only a single spot at a time, they capture images too slowly to track much of the brain’s fast chatter. But specialists are confident that advances in the technology will soon make it possible to see the signals produced by GEVIs at higher speed. “It’s absolutely doable,” Ji says.

If the different approaches can overcome these challenges, scientists have no doubt that voltage imaging will become a standard approach for measuring brain activity. “In the next year or two, we’ll see a lot of papers that have applied voltage sensors and learned about biology,” says Thomas Clandinin, a neurobiologist at Stanford. Some say that the technique might even replace electrodes for questions related to how neurons process and integrate information.

Early-career researchers are particularly optimistic: Hochbaum, who is now a postdoctoral fellow at Harvard Medical School in Boston, says that in the long term, GEVIs will be a go-to tool for studying how different compartments in the cell respond to sub-threshold signals. He plans to use voltage imaging to understand how such signals alter the connection between neurons, a key process in learning. The possibilities are exciting, Hochbaum says, but he’s learnt at least one important lesson from those early days of jumping around the lab in glee after seeing a glow in a microscope: when the experiments work, keep the celebrations to a minimum. ■

Giorgia Guglielmi is a freelance science journalist in Cambridge, Massachusetts.

1. Kralj, J. M., Douglass, A. D., Hochbaum, D. R., MacLaurin, D. & Cohen, A. E. *Nature Meth.* **9**, 90–95 (2012).
2. Piatkevich, K. D. *et al. Nature Chem. Biol.* **14**, 352–360 (2018).
3. Siegel, M. S. & Isaacoff, E. Y. *Neuron* **19**, 735–741 (1997).
4. Xu, Y., Zou, P. & Cohen, A. E. *Curr. Opin. Chem. Biol.* **39**, 1–10 (2017).
5. Yang, H. H. *et al. Cell* **166**, 245–257 (2016).
6. Gong, Y. *et al. Science* **350**, 1361–1366 (2015).
7. Adam, Y. *et al. Preprint at bioRxiv* <https://doi.org/10.1101/281618> (2018).
8. Chien, M.-P. *et al. Preprint at bioRxiv* <https://doi.org/10.1101/211946> (2017).

COMMENT

PHYSICS The struggle for the soul of solid-state science **p.306**



HISTORY Polish team paved the way for Turing to crack Enigma **p.307**

BIODIVERSITY Stakeholders in international panel rise up and respond **p.309**

NATURAL CAPITAL Guidelines, respect and time can reconcile diverse views **p.309**

ACTION PRESS/REX/SHUTTERSTOCK



A pilot project in Spremberg, Germany, aims to capture carbon dioxide released from power stations.

Weigh the ethics of plans to mop up carbon dioxide

Pinning climate hopes on negative emissions technologies is dangerous and demands reflection on the social aspects, warn **Dominic Lenzi** and colleagues.

In October, the Intergovernmental Panel on Climate Change (IPCC) will release a special report on keeping global temperature rise within 1.5°C of pre-industrial levels. Governments requested the report at the 2015 Paris climate conference. Policymakers want to know what further steps would be needed to stay well within the 2°C threshold, above which the risks of climate change become more dangerous.

The IPCC report will confirm an open secret: in the light of growing emissions,

targets for mitigating climate change increasingly depend on 'negative emissions technologies' that suck carbon dioxide out of the atmosphere. Staying within 2°C could mean extracting billions of tonnes of CO₂ this century.

Atmospheric carbon — captured after burning biofuels, for instance — could be locked in the ground or sea for thousands of years. Forests and soils could be managed to store more carbon. Or more-speculative means that are still in the realm of basic

research could be used¹. Examples include fertilizing the oceans with iron to enhance phytoplankton growth, increasing the weathering of minerals or developing devices that remove CO₂ directly from the air.

The vast scale at which such technologies would need to be implemented raises ethical concerns. For example, growing more biomass to burn as fuel would take land away from food production and use water for irrigation². Famines, civil unrest and damage to biodiversity could follow³. ►

► Seeding the oceans with iron could undermine marine ecosystems. Covering an area twice the size of the United States with crushed silicate stones to enhance weathering would affect communities, agriculture and ecosystems.

Yet there has been no systematic evaluation of the ethics of carbon removal methods by the climate assessment community or professional philosophers. The IPCC's latest review (its fifth assessment report) included a chapter on ethics⁴, setting out concepts of responsibility, justice and welfare. But it did not dwell much on negative emissions technologies, nor did other chapters consider ethics. Carbon removal methods must be ethically evaluated in the context of climate policy pathways.

The key question is, which pathways are most compatible with human rights, sustainable development and environmental protection? The stakes are high. Negative emissions technologies could be a valuable way to avoid dangerous climate change. But they might become an unjust gamble that uses future generations as collateral⁵.

MISSING ETHICS

Why has this aspect of negative emissions been overlooked? Ethicists neglect the science; modellers neglect the ethics. Geo-engineering debates have been dominated by solar-radiation management — altering the reflectivity of the whole atmosphere seems more dystopian than growing forests or storing carbon. Early studies suggested that negative emissions technologies were largely benign⁶. Growing dependence on negative emissions increases the risks, but most ethicists have not noted this shift.

Philosophical discussions of climate change revolve around abstract principles: the 'common but differentiated responsibilities of states' to fund mitigation and adaptation, whether the polluter pays and who has the ability to pay. The debates do not consider particular policy pathways, telling us little about what a just future would look like or how to achieve it. Without interrogating mitigation pathways, ethics will be of little use for policy assessment.

Ethicists need a better understanding of climate-mitigation research. The vast scales over which negative emissions technologies would be unleashed are difficult to grasp. Even the climate stabilization target isn't settled. It seems obvious that lower temperatures are ethically preferable. But getting negative emissions wrong also raises risks. Keeping within 1.5 °C of warming could cause side effects that are as bad as those in a world that is 2 °C warmer — such as through environmental damage caused by ramping up mineral mining, or cutting down the rest of the Amazon rainforest for biofuels.

VALUES IN DISGUISE

Meanwhile, modellers inevitably make value-laden assumptions in charting different policy pathways, including the range of options being considered, such as rapid technological development or nuclear energy. Assumptions also include the political, economic and demographic stories behind

them, such as steady population growth or declining international cooperation. For example, the IPCC included negative emissions technologies in its 'default' technology mix, even though some of these solutions might never be viably scaled up⁷.

A lack of transparency and ethical discussion has three consequences. First, policymakers have false expectations. This is the 'moral hazard' worry: if politicians and advisers think it is acceptable to emit carbon now and claw it back later, they might take more risks and obstruct mitigation in the real world⁸. For example, in IPCC scenarios with CO₂ retrieval, emissions from fossil fuels and industry can remain as high as 32 gigatonnes of CO₂ in 2030 (see 'Three-fold folly', top left panel). Without CO₂ removal, emissions would have to be reduced to 23 gigatonnes of CO₂ by 2030 — a difference almost equivalent to China's emissions each year since 2008.

Second, designing climate policy around technologies that might never scale up is risky⁷. A typical 2 °C climate scenario requires the funding, construction and operation of as many as 16,000 plants that combine biomass burning with carbon capture and storage by 2050. Today there are three demonstration projects (see 'Three-fold folly', top right panel). If the bet fails, future generations will face a carbon overdraft and warming that is greater than 2 °C (ref. 3).

Third, implementing negative emissions at the scales envisaged is ambitious, to say the least⁵. According to many models, this would mean managing an artificial carbon sink that is larger than the entire land sink today (see 'Three-fold folly', lower panel). Assessments also must not ignore a host of potential feedback mechanisms and tipping points that are poorly understood, such as whether temperature overshoot might trigger permafrost melting⁹.

DRAW ON DIVERSE VIEWS

A cultural change is required across the climate-change community. As for human-participants research in bioscience, ethicists need to be involved from the outset in developing, modelling and evaluating scenarios for reducing emissions. The IPCC should integrate the perspectives of these experts across all chapters of its reports. There are no plans to do this within the sixth IPCC review currently in preparation.

To broaden the range of considerations included, we would like to see ethicists, modellers and social scientists, governments and civil-society groups collaborate on climate-mitigation assessments. Drawing on divergent viewpoints and criteria, they should map the various implications of alternative policy pathways¹⁰. Organizations such as the Integrated Assessment Modeling Consortium should openly discuss ethical assumptions built into models. This might help to avoid misleading or opaque choices



A palm-oil plantation in the Democratic Republic of the Congo encroaches on nearby rainforest areas.

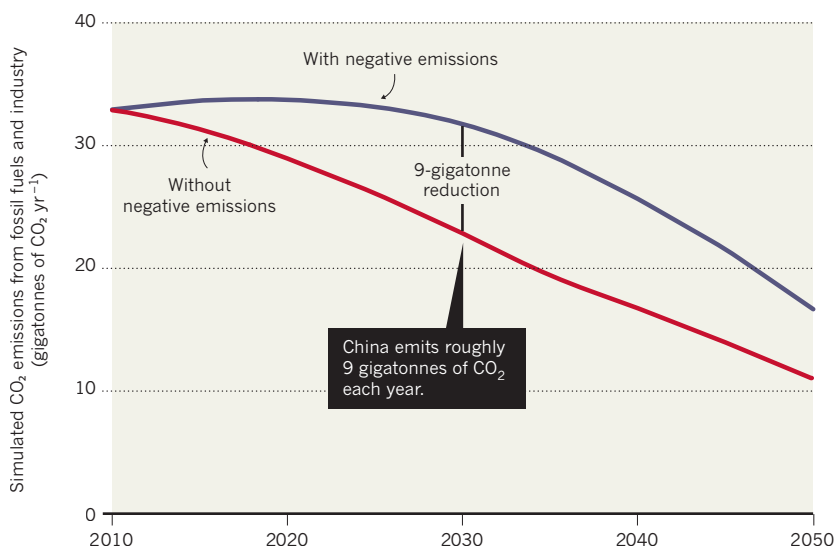
DANIEL BELTRÁ/GREENPEACE

THREE-FOLD FOLLY

Technologies that capture carbon dioxide on a planetary scale might help to avert dangerous levels of climate warming, but they are risky.

COULD DELAY CUTS

Policymakers and industry could delay the reduction of emissions in the belief that these can be clawed back later with negative emissions.



REQUIRES STEEP SCALE-UP

Designing climate policy around technologies that might never sufficiently scale up is a gamble.

□ = 100 biomass power plants with carbon capture and storage

Future generations would bear the burden of failure to scale up negative emissions.

3 demonstration plants exist today

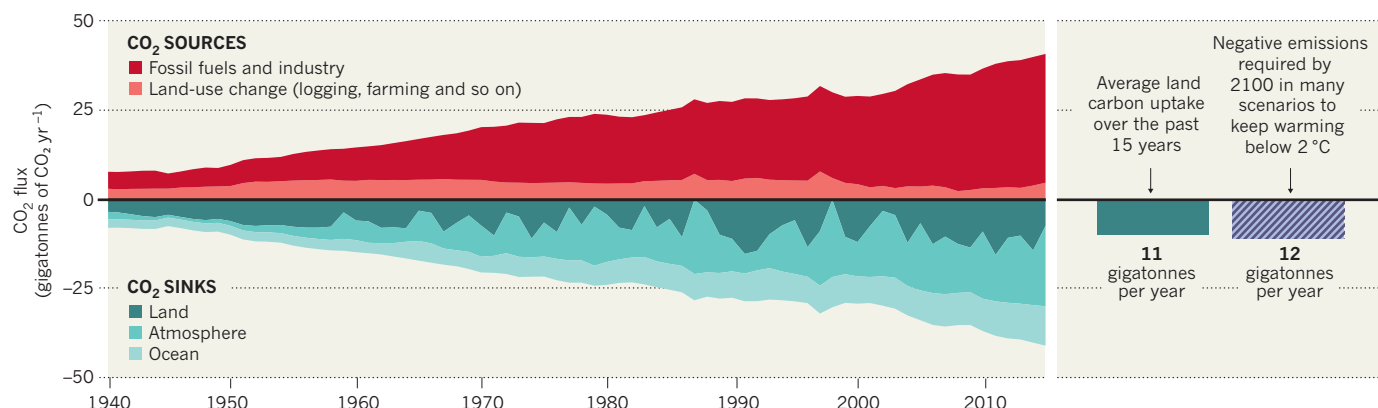
700

5,000

16,000

DEMANDS UNPRECEDENTED SINK

The scale of negative emissions required in many scenarios would mean controlling a massive carbon sink (purple bar) — larger than the entire current natural land sink.



being made at the design stage. For example, lifestyle changes such as meat-free diets or avoidance of aeroplane travel have been absent until recently from scenarios, leading to an imbalanced representation of options¹¹.

Jointly assessing the desirability of alternative futures against ethical principles and the policy goals underlying sustainable development would facilitate critical reflection on negative emissions. Funding bodies such as the European Commission, Future Earth, the US National Science Foundation and other supporters of interdisciplinary research must integrate ethical and social analyses with climate scenario modelling and policy evaluation.

How else can we debate the sort of future we want? ■

Dominic Lenzi, William F. Lamb and Jérôme Hilaire are researchers at the

Mercator Research Institute on Global Commons and Climate Change, Berlin, Germany. Martin Kowarsch is head of scientific assessments, ethics and public policy and Jan C. Minx is head of applied sustainability science at the Mercator Research Institute. J. H. is also at the Potsdam Institute for Climate Impact Research, Germany; and J. C. M. is also at the School of Earth and Environment, University of Leeds, UK.
e-mail: lenzi@mcc-berlin.net

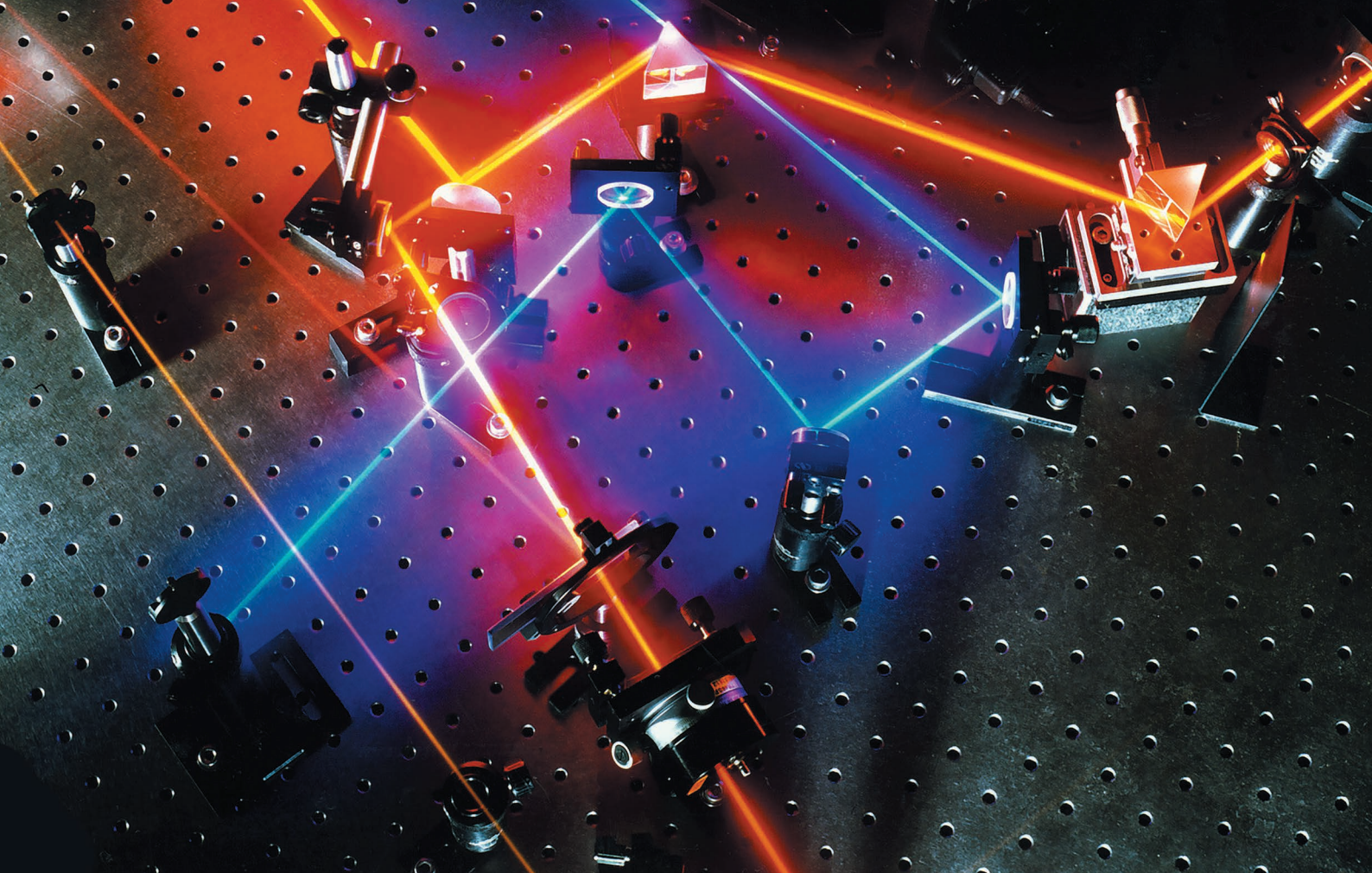
1. Nemet, G. F. *et al. Environ. Res. Lett.* **13**, 063003 (2018).
2. Creutzig, F. *et al. Glob. Change Biol. Bioenergy* **7**, 916–944 (2015).
3. Shue, H. J. *Hum. Rights Environ.* **8**, 203–216 (2017).
4. Kolstad, C. *et al. in Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate*

Change (eds Edenhofer, O. *et al.*) Ch. 3, 207–282 (Cambridge Univ. Press, 2014).

5. Lenzi, D. *Glob. Sustain.* **1**, e7 (2018).
6. Shepherd, J. *et al. Geoengineering the Climate: Science, Governance and Uncertainty* (Royal Soc., 2009).
7. Fuss, S. *et al. Nature Clim. Change* **4**, 850–853 (2014).
8. Minx, J. C. *et al. Environ. Res. Lett.* **13**, 063001 (2018).
9. Fuss, S. *et al. Environ. Res. Lett.* **13**, 063002 (2018).
10. Kowarsch, M. *et al. Nature Clim. Change* **7**, 379–382 (2017).
11. Creutzig, F. *et al. Nature Clim. Change* **8**, 260–263 (2018).

CORRECTION

The timeline in the Comment 'Publish peer reviews' (*Nature* **560**, 545–547; 2018) erroneously stated that peer review began to be published at *The EMBO Journal* in 2010. It was, in fact, in 2009.



Femtosecond laser systems, which emit ultrashort optical pulses, are used to probe fundamental properties of solid-state materials.

PHYSICS

A history of substance

Michael Gordin applauds a study tracing 70 tumultuous years of solid-state physics.

What is “physics”? From the birth of the nuclear age at the end of the Second World War, physics has often been portrayed as the quest to penetrate the atom; to divine the secrets of the subnuclear realm of mesons and quarks with ever more impressive accelerators and ever more gargantuan particle detectors. High-energy physics was the glamorous stuff that attracted Nobel prizes and lavish press coverage. Most books on the history of physics consider it the backbone of the field.

Studying elementary particles is not, however, what most physicists do. By almost all metrics — PhD degrees, articles in flagship journals, memberships in professional societies — the majority were not, and are not, high-energy physicists. Instead, they plough the furrows of what was once known as solid-state physics — better known since the 1970s as condensed-matter physics. This is the science that brought us superconductivity, superfluidity, magnetic memory, liquid-crystal displays and more.

This is the physics that the science historian

Joseph Martin presents in *Solid State Insurrection* — but his focus is not those landmarks. For him, “physics is what physicists decide it is”. This is not some slogan of radical relativism. It is a recognition that physics is a profession, and it is the business of professional groups to police their boundaries.

In the United States, the motivation of much of that policing was access to funding: high-energy physicists got loads of it from the government, and solid-state physicists were shunted to industry. For decades, condensation blossomed on one side and resentment festered on the other. Matters came to a head in a decision by Congress to cancel what would have been the crown jewel of US



Solid State Insurrection: How the Science of Substance Made American Physics Matter

JOSEPH D. MARTIN
University of Pittsburgh Press (2018)

high-energy physics, the Superconducting Super Collider (SSC), in 1993. That cancellation was influenced by criticisms before congressional committees from eminent condensed-matter specialists, such as Nobel laureate Philip Anderson of Bell Labs.

This dispute was about more than resources. Already in the 1970s, solid-state physicists such as Anderson and Alvin Weinberg had articulated an alternative vision of the science of physics. Particle physicists justified themselves through a commitment to “pure science” that dated back to the origins of the American Physical Society (APS) in the late nineteenth century. Because high-energy physics probed the smallest constituents of matter, a reductionist would say that such physics was the most “fundamental”. Anderson disagreed. As Martin explains in an excellent chapter, for Anderson “fundamental physics” was about ferromagnets as well as about quarks.

Then came the SSC. “The original Star Wars trilogy tells the story of a ragtag band of misfits, many of whom are adept at

manipulating a force pervading in everyday matter, who ally to mount an insurrection against the established order and help destroy a giant, partially built beam machine,” writes Martin. The trajectory of US solid-state physics, he notes, “followed much the same plot”. Although he concedes that the SSC was more drastically affected by the end of the cold war than by intradisciplinary critique, there is no doubt where Martin’s sympathies lie.

He devotes most of his book to a detailed reconstruction of the intense struggle, half a century earlier, for recognition by solid-state physicists against the leadership of the APS, which was itself frustrated and challenged by the rapid growth in their ranks during the 1940s. Physicists who worked on metals, ceramics and other domains straddling fundamental and applied physics wanted representation at APS meetings, leading to the creation of the Division of Solid State Physics in 1947. The institutional gerrymandering had significant implications for the APS, especially for its flagship journal, *Physical Review*. (Publishing is a fascinating leitmotif in Martin’s account.)

This organizational innovation was achieved only after substantial resistance from some APS stalwarts, who perceived the purity of their ranks as becoming sullied by industrial scientists. The stalwarts included Harvard University’s John Van Vleck, even though he had trained many of the leaders of the next generation, including Anderson. Van Vleck’s objections were littered with political language: he protested against the “Balkanization” of the APS, and he thought the solid-state division was a “new-deal-bureaucratic” scheme that ought to be resisted. The conservative Van Vleck was unhappy about the direction that the United States — and with it physics — was going.

This raises a broader point about Martin’s engaging book: the politics in it are exclusive to the profession. He keeps his gaze tightly trained on physicists as they define physics to each other. The Vietnam War (and scientific work in support of it), anti-Communism, civil rights and other political fault lines — which affected physicists no less than other citizens — are mentioned in passing, if at all. Yet they must have mattered. Physics is defined not just by what physicists decide it is, but by what the broader society will (or won’t) support. That decision is made within the halls of the APS, but also in those of Congress. ■

Michael D. Gordin is *Rosengarten Professor of Modern and Contemporary History at Princeton University in New Jersey*.
email: mgordin@princeton.edu

Forgotten heroes of the Enigma story

Joanne Baker enjoys a tale of the Polish cryptographers who paved the way for Alan Turing’s wartime feats.

Alan Turing’s crucial unscrambling of German messages in the Second World War was a tour de force of codebreaking. From 1940 onwards, Turing and his team engineered hundreds of electronic machines, dubbed bombs, which decrypted the thousands of missives sent by enemy commanders each day to guide their soldiers. This deluge of knowledge shortened the war. Bletchley Park, UK — the secret centre where it all happened — rightly gained its place in history. But as with all breakthroughs, many more people laid the foundations.

In his book *X, Y & Z*, Dermot Turing, the great mathematician’s nephew, tells the gripping story of a band of Polish mathematicians who worked out much about how German Enigma encoding machines operated, years before Alan Turing did. The Poles shared their secrets with French and British intelligence services before and during the Second World War — the letters X, Y and Z were shorthand for the French, British and Polish codebreaking teams, respectively.

The author’s research is painstaking. After the war, military documents were scattered across Europe, and key French records were declassified only in 2016. Many original Polish papers were destroyed, but



X, Y & Z: The Real Story of How Enigma Was Broken

DERMOT TURING
The History Press
(2018)

the mathematicians’ families have shared personal letters. Turing unearths a remarkable tale of intellect, bravery and camaraderie that reads like a nail-biting spy novel.

Polish skills in cryptography and radio engineering came together during the 1920 Russo-Polish War. Signallers decoded a telegram from Red Army military commander Joseph Stalin, which indicated that an attack on Warsaw was imminent. Jamming the Russians’ radio communications bought enough time to secure and save the city. Maksymilian Ciężki and Antoni Palluth were among those signallers. After the 1920 conflict, Ciężki became leader of a radio-intelligence unit. Palluth set up a business making electronic equipment, including radios the size of a credit card for Polish secret agents.

In 1926, the German navy began to send messages that were scrambled in a more random way, making them almost impossible to decipher. They were encoded using the typewriter-like Enigma machine. The keyboard was wired so that typing one letter lit up a different one in a set of bulbs on top. Rotors altered the path of the electric circuit with every keystroke. The machines were commercially available, but modified for German military use. Without knowing the precise setting of a machine, there was no way to unpick the code.

The book tells how Ciężki hired a group of mathematics students to crack the problem. They worked quietly in basements and in a bunker deep in the woods. Marian Rejewski, an alumnus of Poznań University in Poland, was one of them. At the helm was Gwido Langer, a Pole who had worked in radio intelligence for the Austrian army.

Meanwhile, in France, Gustave Bertrand headed the equivalent unit. The French had a more conventional approach to gathering information: good agents, clandestine meetings and generous pay-offs. Bertrand managed two formidable spies. Rudolf ▶



Mathematician Marian Rejewski in 1942.



ANNA ZYGALSKA-CANNON

Polish cryptographers, including Maksymilian Ciężki (seventh from left) and Gwido Langer (centre back, head just seen), pictured in southern France in 1941.

► Stallmann — code name Rex — was a German card-sharp who had posed as a baron to fleece casino-goers; he picked up languages and people with ease. Rex recruited Hans-Thilo Schmidt, or Agent Asche, whose brother was a colonel in the German army. Schmidt supplied cases of military documents to the French, which Rex received and Bertrand and his colleagues photographed in hotel bathrooms.

Bertrand built up a network for sharing intelligence, including with Poland and the United Kingdom. In 1931, he agreed to supply Langer with German military documents if the Poles would pass back decrypted German messages. One of those documents, passed on by Schmidt, was a manual for Enigma.

Langer, Ciężki and Rejewski leapt on it. They discovered that a panel added to the front of the machine altered the settings, although they still could not tell how the device was wired. They set about collecting coded messages and applied their wits to find clues. Sometimes, the senders made telling mistakes. The German soldiers might use simple sets of three letters, such as QQQ, to broadcast the settings to the receiver. Occasionally, the messages could be guessed: for instance, they often said *maschine defekt*.

By 1936, in the run-up to war, the German military was tightening its communications. In October that year, the senders began to reset the Enigma machines daily. Dermot Turing credits another

Polish mathematician, Jerzy Różycki, with realizing that this altered the frequency of letters, revealing extra information. The team developed tools to work through the hundreds of permutations, including punched cards and a mechanical device with rotors that mimicked Enigma, which, for uncertain reasons, the team called a *bomba*. Both concepts were later used and developed by Alan Turing.

Bertrand fed this information back to Britain's codebreakers, who come across in the book as humorous but aloof. They called the French dispatches — stamped *TRÈS SECRET* in red — scarlet pimpernels. In late July 1939, just over a month before the German army marched into Poland, Bertrand arranged for the respected British cryptologist Dillwyn 'Dilly' Knox (who was already working on Enigma at Bletchley) to meet Langer's team near Warsaw. The Poles wanted to pass on their knowledge. Initially angry that they had beaten him to it, Knox later sent the Poles a silk scarf printed with a horse-racing scene to concede that they had won.

The British immediately ramped up codebreaking efforts at Bletchley Park; within a few months, Alan Turing had re-engineered the bombes to work more quickly.

“Ciężki hired a group of mathematics students who worked in basements and in a bunker deep in the woods.”

The Polish insights saved him a year of work.

When war broke out, the Polish radio-intelligence unit was wound up. The codebreakers buried their notes and machines and fled. Some ended up in Algeria, the rest in France working for Bertrand, who had set up a radio-intelligence group in a chateau in southern France. In breathtaking passages, the book reveals how most of the Polish codebreakers made it through the war, tapping into the French Resistance and dodging Germany's military-intelligence services and secret police when France became occupied. As valuable assets, the codebreakers were not allowed to fight. Touching photographs in the book show them joking with each other and spending time with girlfriends amid the devastation.

Eventually, Ciężki and Langer were arrested and interned in the Sudetenland (now part of the Czech Republic). After the war, they settled in Scotland. Palluth was killed in Germany in 1945, when an aeroplane factory in which he was working at Sachsenhausen concentration camp was bombed. Bertrand artfully played all sides. Avoiding becoming a double agent, he ended up a general. In 1972, he wrote a popular French book about Enigma, and so the story of X, Y and Z and Bletchley began to seep out.

Dermot Turing's vivid and moving account sets the record straight. ■

Joanne Baker is a Senior Comment editor at Nature in London.

Correspondence

Biodiversity: ideas need time to mature

Disagreements over the values of biodiversity are not a problem caused by the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) — nor are they a threat to its success (see *Nature* **560**, 409; 2018). Such debates are grist to the mill of innovation for environmental governance.

People value the natural world in different ways. This is reflected in the ‘ecosystem services’ concept developed through the Millennium Ecosystem Assessment and in the ‘nature’s contributions to people’ approach developed through IPBES. Although both global bodies have sought ways to represent diversity in their assessments, resolving or reducing diverse values has never been a stated role.

The IPBES leadership should therefore not be too hasty in seeking consensus on a single approach to representing values. The IPBES process is still aiming to improve its engagement with the humanities and social sciences (A. Larigauderie *et al.* *Nature* **532**, 313; 2016), and its methodological assessment of biodiversity values started only this year. Both initiatives will catalyse new thinking. Influential ideas take time to mature.

IPBES has committed to move away from focusing solely on scientific assessment. This could achieve something more powerful than scientific consensus for biodiversity, namely a greater understanding of the terms through which humans and nature relate to one another at and across different scales.

Jasper Montana *University of Sheffield, Sheffield, UK.*
j.montana@sheffield.ac.uk

Biodiversity: guide reconciles views

As an ecosystem-services researcher and lead author

of a guide on the values of nature approved by IPBES, I am saddened by the perceived conflict in the biodiversity community over the ‘ecosystem services’ and ‘nature’s contributions to people’ approaches to biodiversity valuation (see *Nature* **560**, 423–425; 2018). Rather than being in competition, they are mutually reinforcing.

The IPBES guide (see go.nature.com/2cna2zn) makes it clear that both concepts are fully integrated into the IPBES approach. Since the guide was released, IPBES has embraced ecosystem services as one of many world views that capture how humans perceive their relationship with nature, alongside those pertaining to individuals and cultures whose conceptualization of nature leaves little room for the human–nature dualism.

Conservation is up against powerful and organized forces. Economic arguments such as avoided costs and jobs generation can influence pro-conservation decisions, as can factors such as health or indigenous and local knowledge. When former US president Barack Obama launched the US climate-change plan in 2015, he focused on the number of childhood asthma cases it would reduce, rather than on its potential economic benefits. The priority is to use all the arguments available to mobilize society’s attention.

Bernardo B. N. Strassburg *International Institute for Sustainability and Pontifical Catholic University, Rio de Janeiro, Brazil.*
b.strassburg@iis-rio.org

Biodiversity: united by a common goal

You call on IPBES to heal “rifts” within the academic community, for example over the concepts and terminology around ‘ecosystem services’ and ‘nature’s contributions to people’ (see *Nature* **560**, 409; 2018). As

chair of IPBES, I stress that both parties are united in their goal to secure a sustainable future for nature and for people.

No matter which conceptual framework is used, the message remains the same: all human societies depend on nature and on the cultural, spiritual, societal and economic benefits it provides. If the natural world continues to degrade, everyone will suffer.

IPBES recognizes that inclusive and constructive discussion is crucial for a better understanding of the global challenges we face, and for reaching a consensus on the key issues. It has therefore always embraced a diversity of views to stimulate and challenge thinking within the academic community.

Including a wide range of stakeholders, knowledge holders and decision-makers from a variety of backgrounds — geographic; gender; and disciplinary, including natural and social sciences, the humanities and people with local and indigenous knowledge — is essential for producing credible and legitimate assessments to inform decision-making.

Already, experts from a wide range of crucial programmes, projects and organizations (including those you mention) are participating in the preparation and rigorous peer review of IPBES assessments.

Robert T. Watson *Potomac, Maryland, USA.*
rtwatson1@gmail.com

Biodiversity: debate underpins change

We strongly object to the tone and content of your discussion on the framing and terminology used to explain the dependence of humans’ wealth, health, happiness and identity on the natural world (see *Nature* **560**, 423–425; 2018). In our view, you magnify the differences of opinion, do not do justice to

the respect held for opposing advocates and oversimplify elements of the conversation.

IPBES is not in competition with the Ecosystem Services Partnership, of which R.d.G. is chair. Their debate centres on which term best serves to protect and sustainably manage the natural world: ‘ecosystem services’ or ‘nature’s contributions to people’. Both organizations have released statements that they stand united against biodiversity loss and ecosystem degradation, and that they will work together to highlight the importance of biodiversity to human well-being. Irrespective of the terminology used, our community is undivided in our knowledge that we fundamentally depend on nature in countless ways.

Debate between peers is central to scientific progress. Including the widest possible range of opinions, expertise, knowledge systems and evidence in that debate is fundamental to the systemic changes that are needed. Together, we are committed to providing all decision-makers with the best possible data and insights to inform better policies, decisions and actions on the health of the natural world we all depend on.

Rudolf de Groot *Ecosystem Services Partnership, Wageningen University, Wageningen, the Netherlands.*

Pavan Sukhdev *TEEB Advisory Board, Geneva, Switzerland.*

Mark Gough *Natural Capital Coalition, London, UK.*
dolf.degroot@wur.nl

CONTRIBUTIONS

Correspondence may be submitted to correspondence@nature.com after consulting the author guidelines and section policies at <http://go.nature.com/cmchno>.

Conflicting evidence for HIV enrichment in CD32⁺ CD4 T cells

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

Descours and colleagues¹ reported a marked enrichment for HIV among CD32a⁺ CD4 T cells in people receiving anti-retroviral therapy (ART). This tiny CD32a⁺ population (0.012% of all blood CD4 T cells) contained a median of 0.56 HIV DNA genomes per cell, and accounted for 26.8–86.3% of HIV DNA in CD4 T cells, thus suggesting that targeting CD32a⁺ CD4 T cells might help to clear HIV reservoirs in vivo. Here, we report our unsuccessful attempts to confirm these findings. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0496-1> (2018).

We first used fluorescence-activated cell sorting (FACS) to sort CD4 T cells with high, intermediate and low levels of CD32 staining (CD32^{hi}, CD32^{int} and CD32^{lo}, respectively) from 10 individuals with chronic HIV infection who were receiving ART (mean duration, 8.8 years; range, 2.7–15). We used cell-staining reagents and gating techniques that matched those used by Descours et al.¹ (see Supplementary Methods and Extended Data Fig. 1). As shown in Fig. 1a, we detected no enrichment for HIV DNA in the CD32^{hi} or CD32^{int} CD4 T cells. Moreover, the CD32^{hi} and CD32^{int} subsets combined accounted for no more than 3% of all HIV DNA copies within circulating CD4 T cells in any of the 10 study participants (Fig. 1b). Post-sort flow cytometry of CD32^{hi} and CD32^{int} populations showed heterogeneous patterns that suggested the formation of T cell–B cell or T cell–monocyte conjugates as the origin of most CD32^{hi} or CD32^{int} CD4 T cells, with separation of these conjugates during sorting (Extended Data Fig. 2).

To rule out the possibility that we had inadvertently obtained false negative results either by excluding HIV-infected, CD32⁺ CD4 T cells using tight light scatter gates or by failing to exclude non-T-cell contaminants, we performed parallel sorts on the same 10 samples using an alternative gating scheme. We used a more inclusive light scatter gate as well as markers for B cells, monocytes, dendritic cells and natural killer cells (Extended Data Fig. 3). Events that were CD3⁺ were separated into fractions that were positive for B cell markers (T–B), positive for one or more other non-CD4-T-cell markers (T–other), or negative for all of these, positive for CD4, and CD32^{hi}, CD32^{int} or CD32^{lo}. Neither CD32^{hi} nor CD32^{int} CD4 T cells were enriched for HIV DNA (Fig. 2a). Similarly, we detected no enrichment for HIV DNA in the T–B and

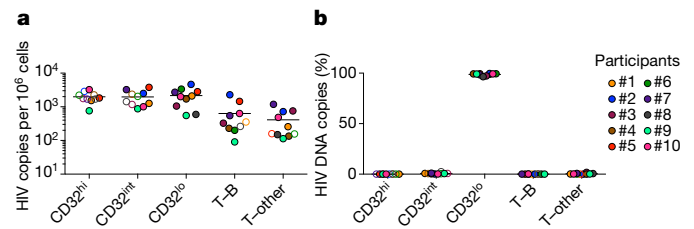


Fig. 2 | Levels of HIV DNA in CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cells, sorted using alternative gating. The samples from Fig. 1 were sorted using alternative gating in which T cells bearing markers of B cells (T–B) or other non-CD4-T-cell lineages (T–other) were first collected in separate tubes. **a**, Copies of HIV DNA per million sorted cells. **b**, Percentages of all HIV DNA copies detected in blood cells that were detected within each subset, calculated by adjusting values in **a** for the relative proportions of these subsets in FACS data.

T–other populations (Fig. 2a). In each of the 10 participants, at least 96% of all HIV DNA copies occurred in conventional CD32^{lo} cells (Fig. 2b). Post-sort flow cytometry suggested that most events bearing both T-cell and non-CD4-T-cell markers again represented cell–cell conjugates, and also showed that most remaining CD32^{hi} CD4 T cells did not reproducibly show a high CD32 signal after sorting (Extended Data Fig. 4). This was in contrast to conventional CD32^{lo} cells, which were uniformly pure in post-sort analyses across participants. In a second group of four individuals whose peripheral blood mononuclear cells (PBMCs) were sorted without previous cryopreservation (Extended Data Fig. 5a), we again found no enrichment for HIV DNA based on CD32 expression (Extended Data Fig. 5b), and also observed that HIV DNA sequences in CD32⁺ CD4 T cells were genetically intermingled with HIV DNA sequences in other CD4 T cells (Extended Data Fig. 5c).

Overall, our studies showed no enrichment for HIV DNA in CD32⁺ CD4 T cells, and also raised questions about the source of the CD32 labelling on these cells. We propose that the CD32 expression associated previously with CD4 T cells could have arisen from adherent non-T-cells or cellular material bearing this marker, and that conjugates containing HIV-infected CD4 T cells could be differentially produced and/or recovered in different laboratories with different sample processing and FACS practices. It is important to acknowledge that these considerations do not explain the discrepancy between the Descours et al. study¹ and ours in the quantities of HIV DNA detected within CD3⁺CD4⁺CD32⁺ sorted material. Nevertheless, we wish to emphasize that our findings do not support targeting CD32 molecules on CD4 T cells in emerging HIV cure strategies.

Methods

Participant recruitment and informed consent were performed under Institutional Review Board (IRB)-approved protocols at the US National Institutes of Health (NIH). For FACS, whole PBMCs were stained with monoclonal antibodies matching those used by Descours et al.¹ (see Supplementary Methods) and sorted on a BD FACSARIA. To evaluate purity, a portion of each population was re-analysed on the flow cytometer after sorting. Virus DNA copies in sorted cells were enumerated by fluorescence-assisted clonal amplification². DNA recovery was quantified by albumin (*ALB*) quantitative PCR. Because the FUN-2 monoclonal antibody used

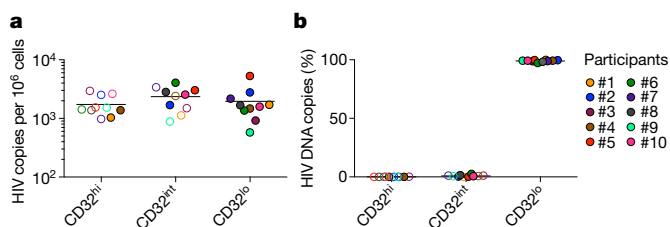


Fig. 1 | Levels of HIV DNA in CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cells, sorted from PBMCs of 10 ART-treated participants, as in Descours et al.¹ **a**, Copies of HIV DNA per million sorted cells. **b**, Percentages of all HIV DNA copies detected in blood CD4 T cells that were detected within each subset, calculated by adjusting values in **a** for the relative proportions of these subsets in FACS data. In all figures, horizontal bars denote median values, and open symbols indicate detection limits for measurements in which HIV DNA was not detected.

BRIEF COMMUNICATIONS ARISING

by Descours et al.¹ and in our study may recognize both CD32a and CD32b, we refer to cells staining with this monoclonal antibody as CD32⁺.

Data availability. All DNA sequences in this manuscript (analysed in Extended Data Fig. 5) have been deposited in GenBank under accession numbers MH080310–MH080572.

Liliana Pérez¹, Jodi Anderson², Jeffrey Chipman³, Ann Thorkelson², Tae-Wook Chun⁴, Susan Moir⁴, Ashley T. Haase⁵, Daniel C. Douek⁶, Timothy W. Schacker^{2,7} & Eli A. Boritz^{1,7*}

¹Virus Persistence and Dynamics Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. ²Division of Infectious Diseases, University of Minnesota, Minneapolis, MN, USA. ³Department of Surgery, University of Minnesota, Minneapolis, MN, USA. ⁴Laboratory of Immunoregulation, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD, USA. ⁵Department of Microbiology and Immunology, University of Minnesota, Minneapolis, MN, USA. ⁶Human Immunology Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. ⁷These authors jointly supervised this work: Timothy W. Schacker, Eli A. Boritz. *e-mail: boritze@mail.nih.gov

Received: 11 October 2017; Accepted: 20 March 2018;

Published online 19 September 2018.

1. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
2. Boritz, E. A. et al. Multiple origins of virus persistence during natural control of HIV infection. *Cell* **166**, 1004–1015 (2016).

Author contributions Data generation and analysis: L.P., J.A., T.W.S. and E.A.B. Study design and oversight: L.P., A.T.H., D.C.D., T.W.S. and E.A.B. Participant cohort and sample management: J.A., J.C., A.T., T.W.C., S.M. and T.W.S. Manuscript preparation: L.P., A.T.H., D.C.D., T.W.S. and E.A.B.

Competing interests Declared none.

Additional information

Extended data accompanies this Comment.

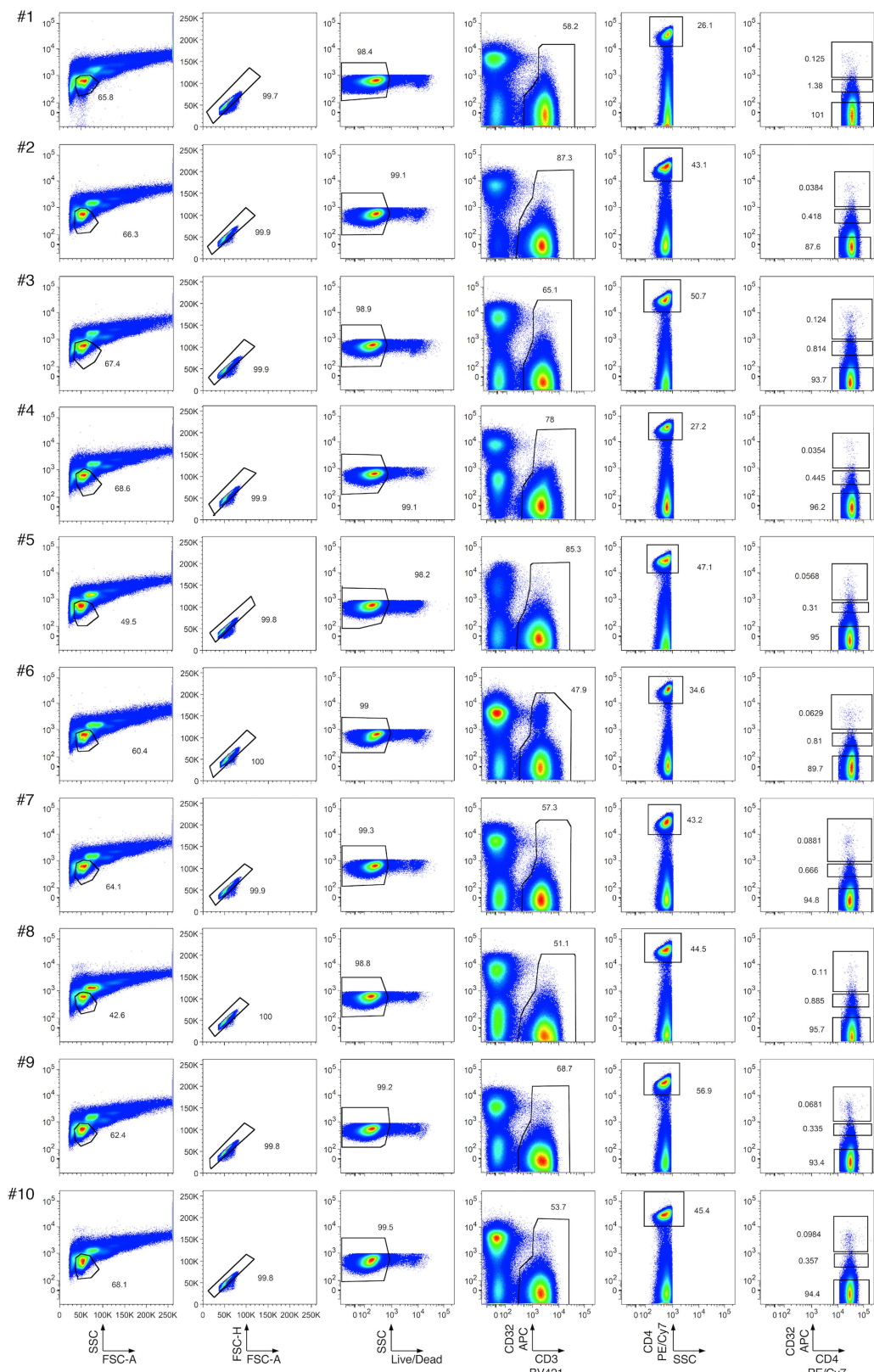
Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to E.A.B.

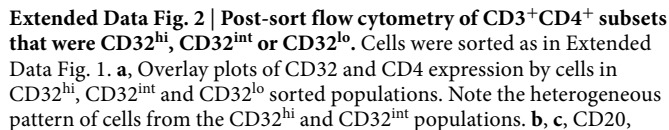
<https://doi.org/10.1038/s41586-018-0493-4>

BRIEF COMMUNICATIONS ARISING



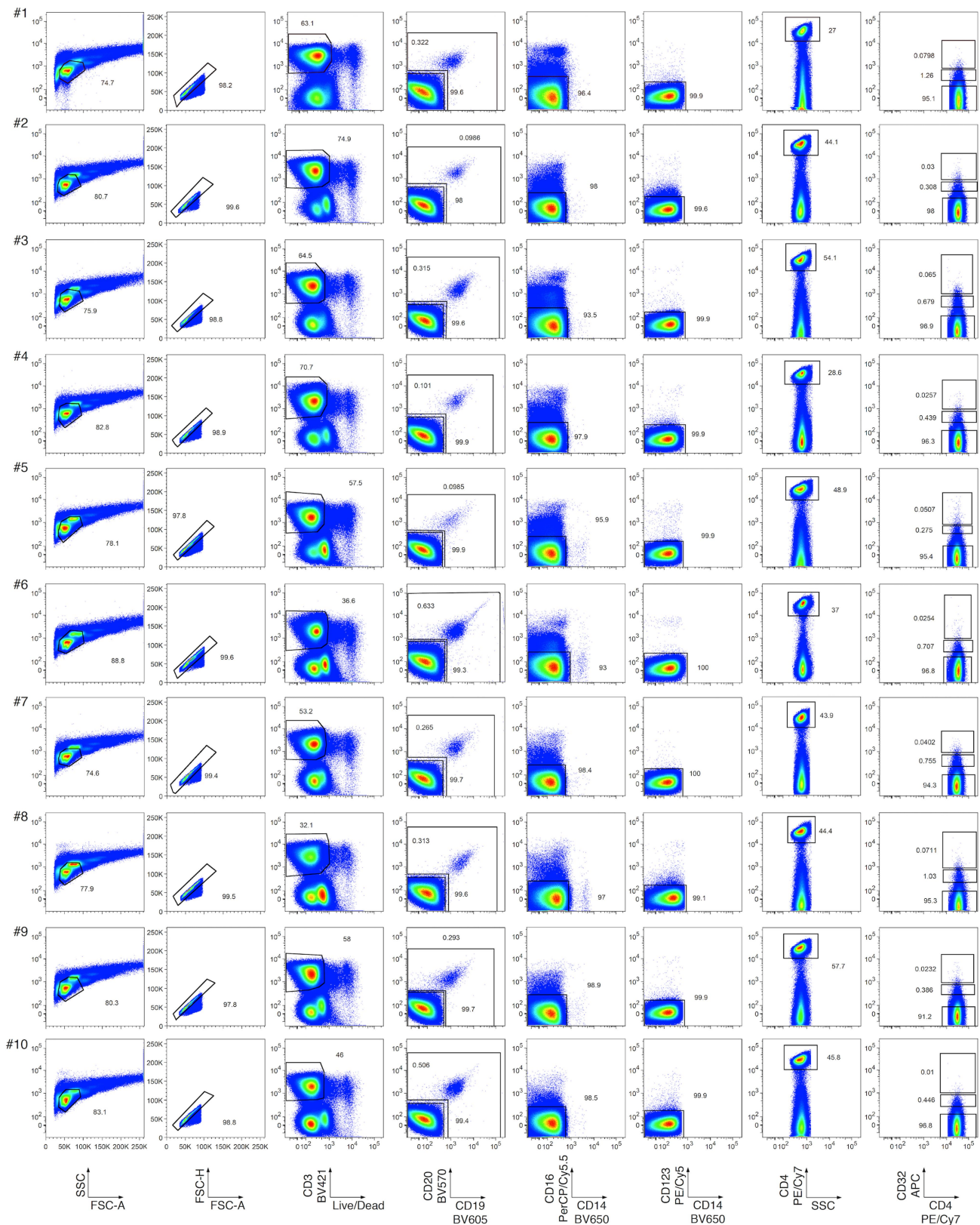
Extended Data Fig. 1 | Flow cytometry of CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations from PBMCs. Single lymphocytes (first two columns) that were viable (third column), CD3⁺ (fourth column), CD4⁺

(fifth column), and CD32^{hi}, CD32^{int} or CD32^{lo} (sixth column) were sorted as described in Descours et al.¹.



E12 | NATURE | VOL 561 | 20 SEPTEMBER 2018

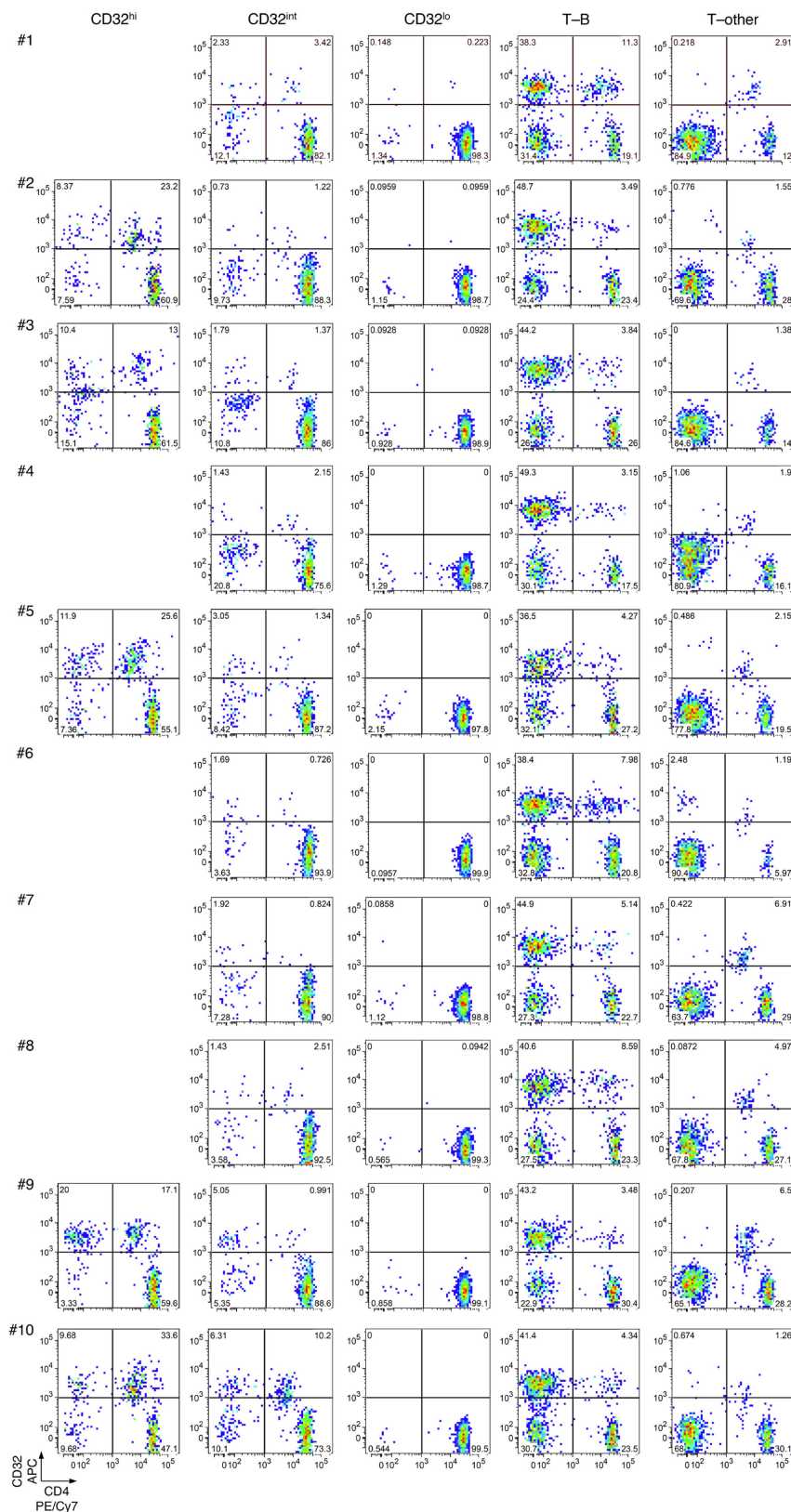
BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 3 | Flow cytometry of PBMCs sorted by alternative gating for CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations, as well as T cell populations bearing markers of B cells (T-B) or other non-CD4-T-cells (T-other). Cells in an inclusive light scatter gate consistent with either small lymphocytes or larger cells (first column) were enriched for single cells (second column). Within these gates, viable CD3⁺ cells

(third column) that were CD19⁻ and CD20⁻ (lower gate, fourth column), CD16⁻ and CD14⁻ (fifth column), CD123⁻ (sixth column), CD4⁺ (seventh column), and CD32^{hi}, CD32^{int} or CD32^{lo} were then collected. Cells that were CD3⁺ and bearing markers of B cells (T-B; upper gate, fourth column) or other non-CD4-T-cells (T-other; combined ungated events from fifth and sixth columns) were also collected in separate tubes.

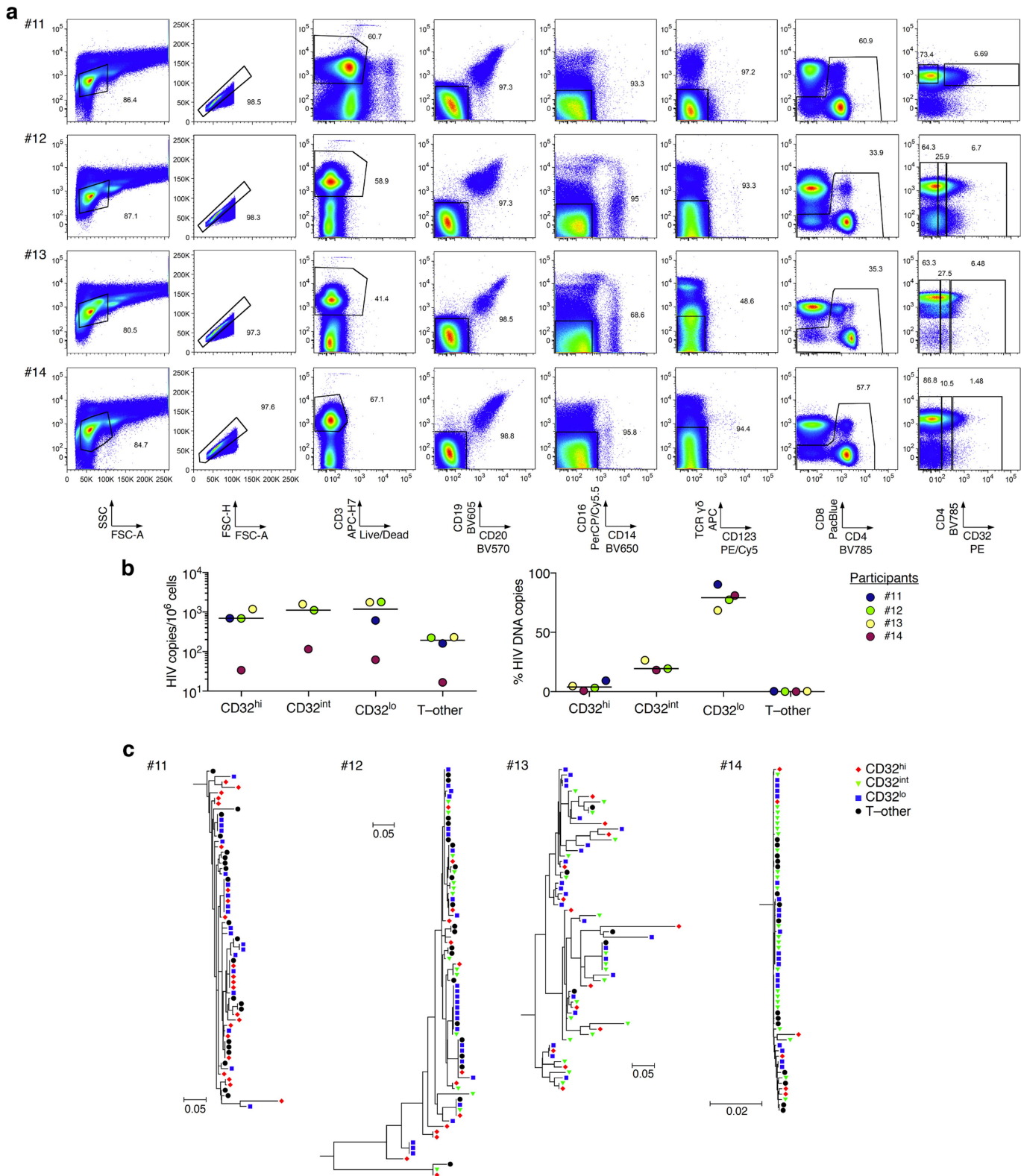
BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 4 | Post-sort flow cytometry of CD32 and CD4 expression by CD32^{hi}, CD32^{int}, CD32^{lo}, T-B and T-other cell subsets. Cells were sorted as in Extended Data Fig. 3. Note the large proportions of all CD32⁺ cells that did not show high CD4 expression after sorting.

Post-sort analyses of CD3⁺CD4⁺CD32^{hi} populations were deferred in cases in which these populations were too small to permit both post-sort analysis and downstream HIV DNA quantification (that is, donors # 1, 4 and 6–8).

BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 5 | See next page for caption.

BRIEF COMMUNICATIONS ARISING

Extended Data Fig. 5 | Flow cytometry, HIV DNA levels, and single-copy HIV DNA sequence analysis from CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations, and from T cells also bearing non-CD4-T-cell markers. **a**, PBMCs from four additional study participants were collected from whole blood by venipuncture with immediate processing (without cryopreservation). The T-other population was collected as a combination of the ungated events from CD19/CD20, CD16/CD14 and $\gamma\delta$ T cell receptor/CD123 exclusion plots (fourth, fifth and sixth columns). **b**, Left, copies of HIV DNA per million cells sorted from four additional study participants as in **a**. Right, percentages of all HIV DNA copies detected in

blood cells deriving from CD32^{hi}, CD32^{int}, CD32^{lo} and T-other subsets, calculated by adjusting values in the left panel for the relative proportions of these subsets determined using FACS data. **c**, Sequences of individual HIV DNA copies were determined by Sanger sequencing of products obtained by fluorescence-assisted clonal amplification, which amplifies a region of the HIV *env* gene. Phylogenetic trees were constructed as described in the Supplementary Methods. All Bonferroni-corrected Slatkin–Maddison *P* values for genetic compartmentalization between any two subsets were greater than 0.05 in all four participants.

The role of CD32 during HIV-1 infection

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

The persistence of latent HIV-1 in resting memory CD4⁺ T cells is a major barrier to a cure, and a biomarker for latently infected cells would be of great scientific and clinical importance^{1–5}. Using an elegant discovery-based approach, Descours et al.⁶ reported that CD32a, an Fcγ receptor not normally expressed on T cells, is a potential biomarker for the HIV-1 reservoir in CD4⁺ T cells⁶. Using a quantitative viral outgrowth assay (qVOA), we show that CD32⁺CD4⁺ T cells do not contain the majority of intact proviruses in the latent reservoir and that the enrichment found by Descours et al.⁶ may in part reflect the use of an ultrasensitive ELISA that does not predict exponential viral outgrowth. Our studies show that CD32 is not a biomarker for the major population of latently infected CD4⁺ T cells. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0496-1> (2018).

If CD32a is a biomarker for latent HIV-1 infection in CD4⁺ T cells, one that is never expressed on CD4⁺ T cells in the absence of HIV-1 infection, then a difference in the frequency of CD4⁺ T cells that express CD32 in HIV-1-infected individuals relative to the frequency in healthy donors is expected. We isolated CD4⁺ T cells from infected and uninfected donors by negative selection and analysed the expression of CD32 and CD4 by flow cytometry. In healthy donors, an average of 0.019% of CD4⁺ T cells was also CD32⁺ (Fig. 1a). This value is not significantly different from levels in HIV-1-infected individuals (Fig. 1a; average 0.011%, $P = 0.1143$) or from values previously reported by Descours et al.⁶ in HIV-1-infected individuals (0.016%, $P = 0.66$). Thus, CD32 does not seem to be a specific biomarker of latently infected CD4⁺ T cells.

To examine whether replication-competent proviruses were present in CD4⁺CD32^{hi} T cells, total CD4⁺ T cells were isolated by negative selection from six HIV-1⁺ individuals that were treated with suppressive anti-retroviral therapy (ART) for at least 6 months (Supplementary Table 1). Freshly isolated cells were stained and sorted to obtain CD4⁺CD32^{hi} and CD4⁺CD32[−] populations, which were analysed in qVOAs⁷ (Fig. 1b, protocol 1). The number of CD4⁺CD32^{hi} cells assayed for each subject is shown in Fig. 1c. On day 14, outgrowth was measured using a standard ELISA for the HIV-1 p24 antigen. CD4⁺CD32^{hi} wells from all subjects were negative for p24 on day 14, and remained negative after an additional week of culture. Conversely, outgrowth was observed in CD4⁺CD32[−] wells from all subjects on both days 14 and 21. The mean infected cell frequency, 1.37 infectious units per million cells (IUPM), was comparable to values previously measured in resting CD4⁺ T cells in several studies (0.03–3.00 IUPM in HIV-1-infected patients⁸, 0.97 IUPM in chronically infected patients⁹) and to values previously measured in the same subjects (mean value 1.33 IUPM) (Fig. 1d, Supplementary Table 2). If the enrichment of proviruses in CD32⁺ cells reported by Descours et al.⁶ was characteristic of replication-competent proviruses, then outgrowth from CD4⁺CD32^{hi} T cells should have been seen (Fig. 1e).

One possible explanation for the discrepancy between our results and those of Descours et al.⁶ is that some latent HIV-1 may be present in a previously undescribed population of CD4⁺ T cells that express CD32 together with other non-T-cell lineage markers. Such cells would be removed during the negative selection used to isolate CD4⁺ T cells. Therefore, we freshly isolated total CD4⁺ cells from infected donors on suppressive ART using two methods: negative selection to remove other lineages, leaving untouched CD4⁺ T cells, and positive selection for

cells expressing CD4 (Fig. 1b, protocol 2). Both CD4⁺ populations were analysed by qVOA. No significant differences were observed in the frequencies of latently infected cells (Fig. 1f). Furthermore, no significant differences in proviral DNA were observed between the purified cell populations (Fig. 1g). Because CD4 is required for HIV-1 entry into the host cell, cell populations obtained via positive selection for CD4 should include every latently infected CD4⁺ T cell. Given that neither the infected cell frequencies nor the levels of proviral DNA differed between the purified cell populations, we conclude that no additional sizable population of latently infected cells was recovered by positive CD4 selection.

In further studies, we used a cell sorting strategy identical to that of Descours et al.⁶ on samples freshly isolated from six subjects receiving ART treatment. Peripheral blood mononuclear cells (PBMCs) isolated from subjects were stained and sorted to obtain CD3⁺CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32[−] cell populations that were tested for latently infected cells by qVOA analysis. The numbers of CD3⁺CD4⁺CD32^{hi} cells assayed for each subject are shown in Fig. 1c and Supplementary Table 3. In addition, total CD4⁺ cells were obtained by staining PBMCs for CD4 and sorting for CD4⁺ cells (Fig. 1b, protocol 3). qVOA results showed that both the CD3⁺CD4⁺CD32[−] and the total CD4⁺ T cell populations had the same infected cell frequencies that were comparable to frequencies measured in other studies¹⁰. However, we observed no outgrowth in CD3⁺CD4⁺CD32^{hi} cultures (Fig. 1h, Supplementary Table 2).

We also analysed CD3⁺CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32[−] cells isolated by the method of Descours et al.⁶ for the presence of proviral DNA by qPCR. We found 89 copies of *gag* per million CD3⁺CD4⁺CD32[−] cells, which is similar to previous measurements in total CD4⁺ T cells¹¹. However, no proviral DNA was detected after DNA extraction from 39,000 CD3⁺CD4⁺CD32^{hi} cells and subsequent qPCR analysis (data not shown). This finding makes it highly unlikely that this cell population is enriched for HIV-1 to a level of more than one provirus copy per cell, as reported by Descours et al.⁶. We caution that the normalization of very low-level HIV-1 DNA measurements from qPCR reactions done with a low number of input cells could artificially produce apparent enrichments in HIV-1 DNA.

In a further attempt to explain the discordant qVOA results obtained in our studies and those of Descours et al.⁶, we tested whether the use of the ultra-sensitive p24 digital ELISA¹² and the low cell input can affect IUPM calculations, leading to erroneous overestimation of latent infection. qVOA culture supernatants were assayed for HIV-1 p24 using the ultrasensitive SIMOA p24 2.0 assay (Quanterix) on days 5, 9, 14 and 21. Using the lower limit of quantification (0.01 pg ml^{−1}) as the cut-off level, we found that two out of three qVOAs containing CD4⁺CD32^{hi} cells tested positive for p24 by this assay, even though the same wells were negative by standard ELISA, which is several orders of magnitude less sensitive (Fig. 2a). Exponential outgrowth is the hallmark of replication-competent viruses. In qVOA cultures of CD4⁺CD32[−] cells, only a fraction of the wells that were positive by SIMOA showed exponential outgrowth as determined by standard ELISA on day 21 (Fig. 2b). Importantly, CD4⁺CD32^{hi} culture wells that tested positive by SIMOA p24 assay showed no exponential outgrowth and had significantly lower levels of p24 (Fig. 2c). It is possible that low positive SIMOA values could reflect an assay artefact or the presence of defective proviruses that are still capable of producing low levels of Gag¹³. A further concern

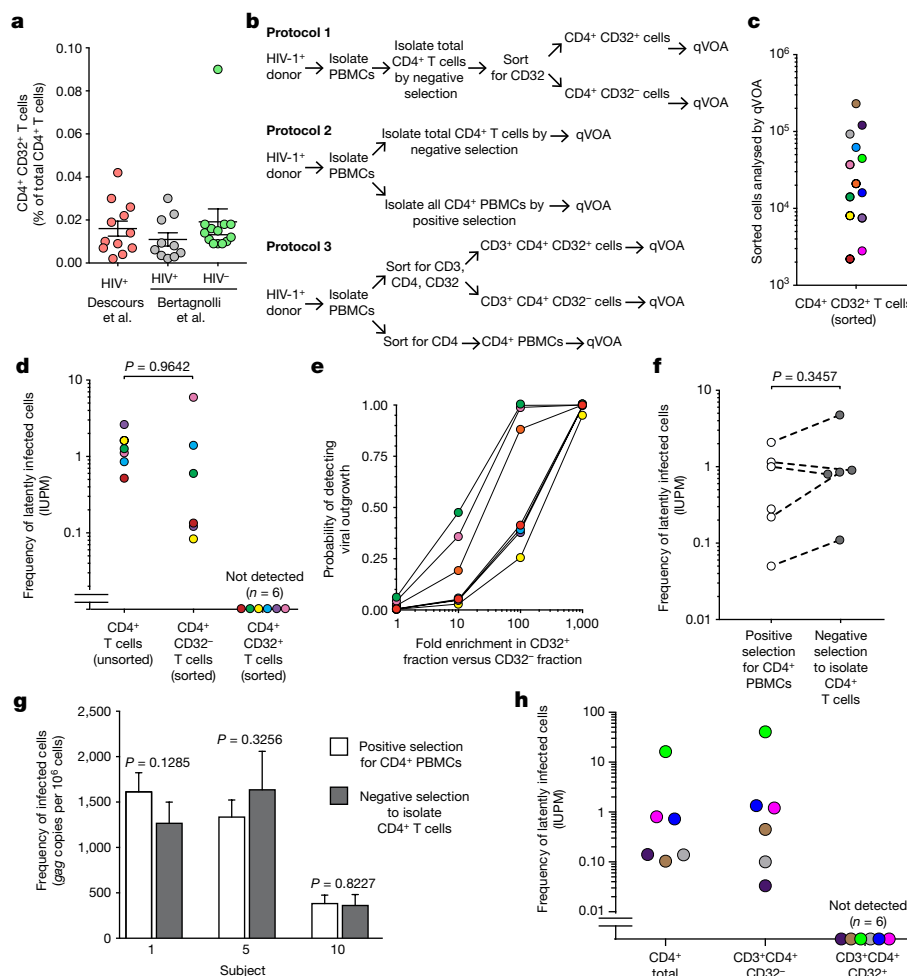


Fig. 1 | Analysis of CD4⁺CD32⁻ and CD4⁺CD32⁺ populations by qVOA and proviral DNA measurements. **a**, Percentage of CD4⁺CD32^{hi} T cells relative to total CD4⁺ T cells in healthy donors and HIV-1-infected donors. Infected donor values were obtained from supplementary table 4 of Descours et al.⁶. LLOQ, lower limit of quantification. **b**, Schematic depicting the three strategies (protocols 1–3) used to obtain different populations of CD4⁺ T cells analysed in qVOAs. **c**, Numbers of sorted CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32^{hi} T cells from each subject analysed in qVOAs. **d**, Frequencies of latently infected cells among CD4⁺CD32^{hi} T cells and CD4⁺CD32⁻ T cells and among total CD4⁺ T cells from the same subjects previously measured in separate experiments. Cells were isolated using protocol 1 (colours correspond to subject values from panel c). **e**, Probability of detecting outgrowth based on measured frequencies of latently infected cells among the CD4⁺CD32⁻ fraction and number of CD4⁺CD32^{hi} cells plated assuming various degrees of enrichment of HIV-1 in CD32^{hi} cells. **f**, Frequencies of latently infected cells measured in qVOAs using positive or negative selection to obtain total CD4⁺ cells (protocol 2; positive selection was accomplished by either sorting or CD4 microbead strategies, with similar results). **g**, Comparison of proviral DNA measurements obtained with qPCR on total CD4⁺ cells purified using positive or negative selection (protocol 2). **h**, Frequencies of latently infected cells among total CD4 cells, and CD3⁺CD4⁺CD32⁻ and CD3⁺CD4⁺CD32^{hi} populations. Cells were isolated using protocol 3 (colours correspond to subject values from panel c).

is that the IUPM calculations are based on cell input, fold dilutions and technical replicates¹⁴, and thus, qVOA analyses performed with very small numbers of sorted CD4⁺CD32^{hi} cells can markedly skew the frequency of cells harbouring replication-competent proviruses (five-fold dilutions from 800 to 1 cell in Descours et al.⁶). When we applied the results obtained with the SIMOA p24 assay, IUPM values ranged from 0 to 3,134 and 554 (patients 4 and 5, respectively; Fig. 2d). As a consequence, when we calculated the ‘fold enrichment’ of IUPM in the CD4⁺CD32^{hi} cells compared to the CD4⁺CD32⁻ cells, we observed a mean fold enrichment of 665 (range 152–1179, from the two patients with positive p24 using SIMOA), similar to what was reported by Descours et al.⁶ (Fig. 2e).

In summary, we find no evidence that CD32 expression indicates the presence of latent HIV-1, and demonstrate that at least a substantial fraction of the HIV-1 latent reservoir is in CD3⁺CD4⁺CD32⁻

T cells. Although no outgrowth could be found in cultures containing CD4⁺CD32^{hi} T cells, viral outgrowth comparable to historical measurements was found in cultures containing CD4⁺CD32⁻ T cells. The use of an ultrasensitive p24 ELISA assay may account for the apparent enrichment observed in culture experiments by Descours et al.⁶. In short, our results have demonstrated that CD32 does not define the HIV-1 reservoir and that future research is needed to identify biomarkers for latently infected cells.

We thank the study participants without whom this research would not be possible. Funding was provided by the US National Institutes of Health (NIH) Martin Delaney I4C, Beat-HIV and DARE Collaboratories by the Johns Hopkins Center for AIDS Research (P30AI094189), by NIH grant 43222, and by the Howard Hughes Medical Institute and the Bill and Melinda Gates Foundation.

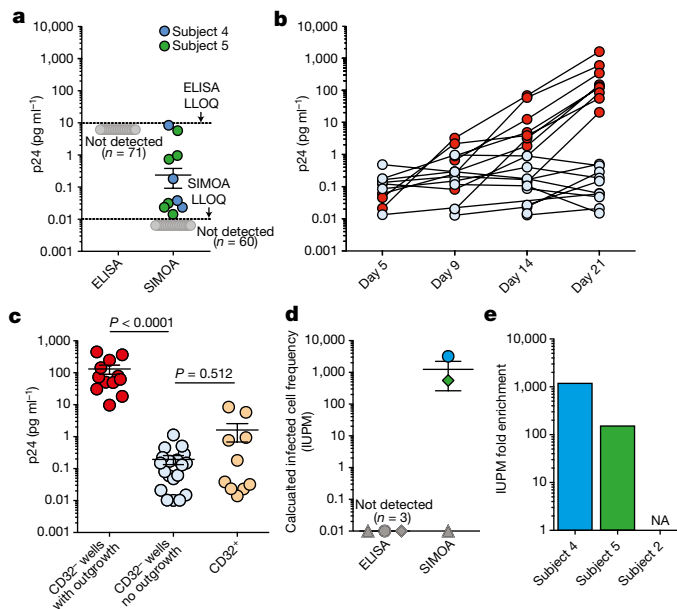


Fig. 2 | Ultrasensitive p24 measurements. **a**, Levels of p24 from CD32⁺ culture wells measured by ELISA and SIMOA (lower limit of quantification: 5–10 pg ml⁻¹ and 0.01 pg ml⁻¹, respectively) (data collected from three subjects, for a total of 71 wells). **b**, Longitudinal levels of p24 measured by SIMOA in individual culture wells in the qVOA for CD32⁻ cells from subject 5, showing wells with and without viral outgrowth (red and blue circles, respectively). **c**, Levels of p24 measured by ELISA in CD32⁻ wells with outgrowth compared with SIMOA measurements in wells with no outgrowth and CD32⁺ wells (data collected from subjects 2, 4 and 5). *P* values were determined with a non-parametric *t*-test. **d**, IUPM calculation based on ELISA and SIMOA analysis. Symbols in dark grey represent values below the limit of detection. **e**, Fold enrichment of IUPM in CD32⁺ cells (from subjects 2, 4 and 5). NA, not applicable.

Methods

qVOAs isolated CD4⁺ T cells using negative depletion and were sorted for CD32⁺ cells (Fig. 1b, protocol 1). To test whether negative depletion was causing a loss of CD32⁺ CD4⁺ T cells, outgrowth and proviral DNA were compared from qVOAs in which CD4⁺ T cells were isolated using positive selection to measurements using negative depletion. Outgrowth measurements and proviral DNA were also measured using the methods described by Descours et al.⁶. Proviral DNA measurements were performed using qPCR¹⁵. HIV-1 p24 values were measured using both a standard ELISA for p24 antigen (Perkin Elmer) and SIMOA (Quanterix). Further details are provided in Supplementary Methods.

Data availability. All data are available from the corresponding author upon reasonable request.

Lynn N. Bertagnoli¹, Jennifer A. White¹, Francesco R. Simonetti¹, Subul A. Beg^{1,2}, Jun Lai^{1,2}, Costin Tomescu³, Alexandra J. Murray¹, Annukka A. R. Antar¹, Hao Zhang⁴, Joseph B. Margolick⁴, Rebecca Hoh⁵, Stephen G. Deeks⁵, Pablo Tebas⁶, Luis J. Montaner³, Robert F. Siliciano^{1,2*}, Gregory M. Laird¹ & Janet D. Siliciano¹

¹Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ²Howard Hughes Medical Institute, Baltimore, MD, USA. ³The Wistar Institute, Philadelphia, PA, USA. ⁴Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. ⁵Division of HIV, Infectious Diseases and Global Medicine, University of California, San Francisco, CA, USA. ⁶Division of Infectious Diseases, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. *e-mail: rsiliciano@jhmi.edu

Received: 29 September 2017; Accepted: 3 April 2018;

Published online 19 September 2018.

1. Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
2. Chun, T. W. et al. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
3. Wong, J. K. et al. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
4. Richman, D. D. et al. The challenge of finding a cure for HIV infection. *Science* **323**, 1304–1307 (2009).
5. Deeks, S. G. et al. Towards an HIV cure: a global scientific strategy. *Nat. Rev. Immunol.* **12**, 607–614 (2012).
6. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
7. Laird, G. M., Rosenbloom, D. I., Lai, J., Siliciano, R. F. & Siliciano, J. D. Measuring the frequency of latent HIV-1 in resting CD4⁺ T cells using a limiting dilution coculture assay. *Methods Mol. Biol.* **1354**, 239–253 (2016).
8. Siliciano, J. D. et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4⁺ T cells. *Nat. Med.* **9**, 727–728 (2003).
9. Eriksson, S. et al. Comparative analysis of measures of viral reservoirs in HIV-1 eradication studies. *PLoS Pathog.* **9**, e1003174 (2013).
10. Crooks, A. M. et al. Precise quantitation of the latent HIV-1 reservoir: implications for eradication strategies. *J. Infect. Dis.* **212**, 1361–1365 (2015).
11. Besson, G. J. et al. HIV-1 DNA decay dynamics in blood during more than a decade of suppressive antiretroviral therapy. *Clin. Infect. Dis.* **59**, 1312–1321 (2014).
12. Passaes, C. P. & Sáez-Cirión, A. HIV cure research: advances and prospects. *Virology* **454–455**, 340–352 (2014).
13. Pollack, R. A. et al. Defective HIV-1 proviruses are expressed and can be recognized by cytotoxic T lymphocytes, which shape the proviral landscape. *Cell Host Microbe* **21**, 494–506.e4 (2017).
14. Rosenbloom, D. I. et al. Designing and interpreting limiting dilution assays: general principles and applications to the latent reservoir for human immunodeficiency virus-1. *Open Forum Infect. Dis.* **2**, ofv123 (2015).
15. Massanella, M., Gianella, S., Lada, S. M., Richman, D. D. & Strain, M. C. Quantification of total and 2-LTR (long terminal repeat) HIV DNA, HIV RNA and herpesvirus DNA in PBMCs. *Bio Protoc.* **5**, e1492 (2015).

Author contributions L.N.B., J.A.W., G.M.L., R.F.S. and J.D.S. designed experiments. S.A.B., C.T. and L.J.M. obtained samples. L.N.B., J.A.W., S.A.B., G.M.L., R.F.S., J.L., A.J.M., A.A.R.A. and J.D.S. performed experiments. R.F.S., H.J. and J.B.M. performed cell sorting. L.N.B., J.A.W., R.F.S., A.J.M., A.A.R.A., R.F.S. and J.D.S. analysed the data and wrote the manuscript.

Competing interests Declared none.

Additional information

Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.F.S.

<https://doi.org/10.1038/s41586-018-0494-3>

Evidence that CD32a does not mark the HIV-1 latent reservoir

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

A recent report by Descours et al.¹ suggests that the cell surface expression of the low affinity Fc receptor CD32a (also known as FcγRIIa) marks the replication-competent HIV-1 reservoir in CD4⁺ T cells from 12 HIV-1-infected participants receiving suppressive anti-retroviral therapy (ART)¹. We have undertaken considerable efforts to replicate these findings using peripheral blood mononuclear cells (PBMCs) from 20 HIV-1-infected, ART-suppressed participants (Extended Data Table 1). We found no evidence to suggest that CD32a marks a CD4⁺ T cell population enriched in either HIV-1 DNA or replication-competent HIV-1 in our study participants. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-s41586-018-0496-1> (2018).

To validate these findings, we adopted the same gating strategy as described by Descours et al.¹ to define CD4⁺ T cell populations (Supplementary Fig. 1a). The CD32 antigen was identified using the same antibody clone (FUN-2) as described by Descours et al.¹. We observed the same CD4⁺ T cell subsets that stained at a high cell surface density of CD32 (CD4⁺CD32^{high}), an intermediate cell surface density of CD32 (CD4⁺CD32^{int}), and a CD4⁺ T cell subset lacking CD32 expression (CD4⁺CD32^{neg}). We obtained frequencies of CD4⁺CD32^{high} T cells that ranged from 0.002% to 0.026%, with a median value (0.012%) that was identical to that reported by Descours et al.¹ (Extended Data Table 2 and Supplementary Fig. 1a). Notably, we confirmed that this same CD4⁺CD32^{high} population is also present in PBMCs isolated from eight healthy donors and exists at similar frequencies to that in HIV-1-infected samples ($P = 0.971$, Extended Data Fig. 1a).

Next, we assessed the amount of replication-competent HIV-1 isolated from the same 20 participants by measuring the infectious unit per million cells (IUPM) in CD4⁺ T cells (range 0.01–37.5, median 0.46). Participant CD4⁺CD32^{high} T cell populations were colour-coded in descending order, and then divided into quartiles that corresponded to the relative frequency of CD4⁺CD32^{high} cells present in these samples (Fig. 1a).

After cytometric sorting of the various CD4⁺CD32 subsets, we quantified HIV-1 DNA in each population (total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high}, Fig. 1b) using droplet digital PCR (ddPCR), as described in the Methods. We found no evidence of HIV-1 DNA enrichment in the CD4⁺CD32^{high} fraction. We observed no significant difference in HIV-1 DNA between any populations and the CD4⁺CD32^{high} T cell population ($P = 0.28$). In fact, levels of HIV-1 DNA in the CD4⁺CD32^{high} T cell subsets isolated from nine participants was at the assay limit of detection (Fig. 1b). After correction for cell input in the CD4⁺CD32^{high} fraction, as estimated DNA values, we saw no evidence for HIV-1 DNA enrichment (open symbols in Extended Data Fig. 1b).

We then compared the relative frequency of the CD4⁺CD32^{high} T cell populations and the viral replicative capacity (IUPM values) per participant, but no relationship between the two parameters was observed (Fig. 1c). All values have been tabulated in Extended Data Table 2.

The HIV-1 reservoir largely resides in quiescent CD4⁺ T cells^{2,3}. Therefore, we sought to confirm the activation status of the CD4⁺ T cell populations by measuring the frequency of the activation markers CD69, CD25 and HLA-DR on CD4⁺ T cell subsets from all

participants. We found that the CD4⁺CD32^{high} T cells were highly activated compared to the CD4⁺CD32^{neg} T cells ($P < 0.0001$). Notably, among the activation markers, HLA-DR was particularly enriched,

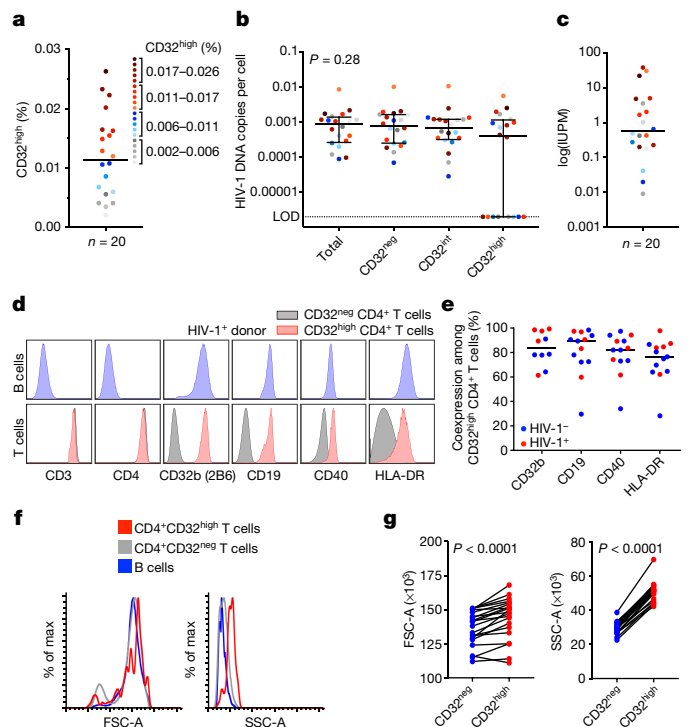


Fig. 1 | CD32-expressing CD4⁺ T cells are not enriched in HIV-1 DNA and express markers of B cell origin. **a–c**, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells from PBMCs of ART-suppressed, HIV-1-infected patients ($n = 20$) were sorted, and HIV-1 DNA was measured by ddPCR. **a**, Dividing the frequency (in percentage) of CD4⁺CD32^{high} T cells from all participants into quartiles, the values are shown as below or above the median. **b**, DNA copies per cell in sorted subsets of total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells are shown, with median and interquartile range (IQR). P value determined by Kruskal–Wallis test. LOD, limit of detection. **c**, IUPM in CD4⁺ T cells of each participant is shown in the colour corresponding to its frequency of CD4⁺CD32^{high} cells in panel **a**. **d**, **e**, CD32^{neg} and CD32^{high} (identified using FUN-2) CD4⁺ T cells from human PBMCs were assessed by flow cytometry for the expression of CD32b (2B6 antibody), CD19, CD40 and HLA-DR and compared to B cells (CD3⁺CD14⁺CD19⁺ lymphocytes). **d**, Representative flow cytometry results per cell antigen levels on B cells (top, blue histograms) and on CD32^{neg} and CD32^{high} CD4⁺ T cells (bottom, grey and red histograms, respectively) from PBMCs from an HIV-1⁺ participant. **e**, Frequency of CD4⁺CD32^{high} T cells staining positive for CD32b (2B6), CD19, CD40 or HLA-DR from HIV-1⁺ ($n = 5$) and HIV-1[−] ($n = 5–8$) human donor PBMC samples. Bars denote median values. **f**, Representative histograms of the FSC-A and SSC-A of B cells and CD32^{neg} and CD32^{high} CD4⁺ T cells from PBMCs of an HIV-1⁺, ART-suppressed participant sorted on a BD FACSAria II. **g**, Comparisons of the median FSC-A and SSC-A values between CD32^{neg} and CD32^{high} CD4⁺ T cell subsets from HIV-1⁺, ART-suppressed participants ($n = 20$). P values were determined using a paired t -test.

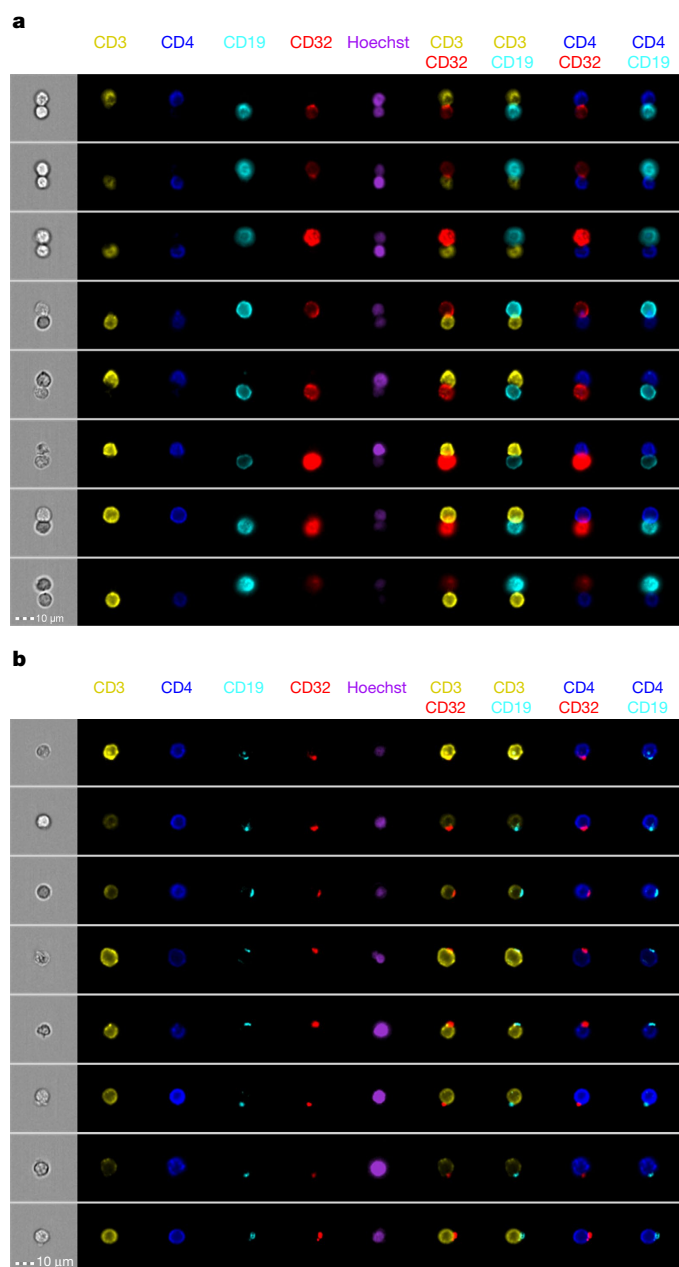


Fig. 2 | Flow cytometry imaging of sorted CD32-expressing T cells. **a**, Representative bright-field and pseudo-colour fluorescence images of T–B cell conjugates found in the CD32^{high} CD4⁺ T cell population sorted from PBMCs from HIV-1⁺, ART-suppressed participants, and imaged using Amnis technology. **b**, Representative images of punctate CD32 staining found on single T cells in the CD32^{high} and CD32^{int} population sorted from HIV-1⁺, ART-suppressed participant PBMCs.

marking approximately 75% of all CD4⁺CD32^{high} cells (median 74%), compared to CD4⁺CD32^{neg} cell populations (median 1.4%) (Extended Data Fig. 1c, $P < 0.0001$).

Two CD32 isoforms (CD32a and CD32b) are known to be expressed among all antigen presenting cells (APCs), but not typically on T cells. Therefore, we sought to exclude any APCs as potential contaminants of flow cytometry sorting. We evaluated the co-expression of lineage markers for all major CD32-bearing cells including monocytes, B cells, dendritic cells, granulocytes and natural killer cells. As expected, all CD32⁺ T cells expressed high amounts of CD3 and CD4 (Fig. 1d). However, we found that most CD4⁺CD32^{high} T cells from HIV-1⁺

patients, and also from healthy donors, co-expressed several B cell markers including CD19, CD40 and HLA-DR (Fig. 1d, e, Extended Data Fig. 2a). Notably, the B cell antigens found on CD4⁺CD32^{high} T cells were present at similar cell-surface densities as detected on bona fide B cells (Fig. 1d, e, Extended Data Fig. 2a).

The demonstration that the CD4⁺CD32^{high} fraction seen in HIV-1⁺ patients was marked with several B cell antigens and was similarly present in naive donors led us to investigate the origin of these B cell markers on a CD4⁺ T cell. Several reports have shown that B cells exclusively express the CD32b isoform⁴. The FUN-2 antibody clone used by Descours et al.¹ cannot distinguish between the CD32a and CD32b isoforms. Therefore, we used the monoclonal antibody clone 2B6 that has been reported to exclusively bind to CD32b^{4,5}. After co-staining PBMCs from HIV-1⁺ and HIV-1⁻ individuals with both the FUN-2 and the 2B6 antibodies, we found that all CD4⁺CD32^{high} T cells were marked only by the CD32b isoform and not by CD32a (Fig. 1d, e, Extended Data Fig. 2b), indicating that B cells are the origin of the CD32b antigen that marks the CD4⁺CD32^{high} T cells.

We sought to confirm this by determining whether *CD32A* (also known as *FCGR2A*) or *CD32B* (*FCGR2B*) mRNA was endogenously produced in the CD4⁺CD32^{high} subsets. After isolating total cellular RNA from various sorted T cell subsets, we used established reverse transcription PCR (RT–PCR) primers and probes that are specific to the CD32a and CD32b isoforms, as described in the Methods. We found that sorted CD4⁺CD32^{high} T cells from four HIV-1-infected participants did not contain detectable levels of the *CD32A* isoform. However, the *CD32B* mRNA isoform was readily detected in CD4⁺CD32^{high} T cells isolated from two out of four HIV-1⁺ patients (Extended Data Fig. 2c). By additional RT–PCR analysis, we detected both *CD3G* and *CD19* transcripts in the same CD4⁺CD32^{high} T population, indicating that the CD32b marking the CD4⁺CD32^{high} T cells may be from B cells expressing cognate CD32b (Extended Data Fig. 2d).

Because this may require cell-to-cell interaction, we performed a back-gating analysis of our flow cytometry data and confirmed that all CD4⁺CD32^{high} populations were identified within single-cell gates (Supplementary Fig. 1b). However, post-hoc analysis comparing the forward and side scatter light pulse area (FSC–A and SSC–A, respectively) values between CD4⁺CD32^{neg} and CD4⁺CD32^{high} T cells showed that the CD4⁺CD32^{high} populations had both a significantly higher FSC–A ($P < 0.0001$) and SSC–A ($P < 0.0001$), suggesting that the CD4⁺CD32^{high} population may consist largely of cell doublets (Fig. 1f, g).

We next used Amnis imaging flow cytometry to visualize the sorted CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} cell populations directly. As expected, the CD4⁺CD32^{neg} and the CD4⁺CD32^{int} cell populations each consisted of more than 99% single cells. However, the CD4⁺CD32^{high} fraction contained a high frequency of cell doublets (mean value 94%) (Extended Data Fig. 3). Of these ‘doublets’, approximately 70% seemed to be coincident doublets, and 30% were conjugates of T and B cells (Fig. 2a and Extended Data Fig. 3b).

We observed no examples in which CD32 staining on T cells was distributed throughout the cell membrane, supporting the idea that the CD32 found in the CD4⁺CD32^{high} population is not the result of endogenous expression from CD4⁺ T cells. Of the instances in which CD32 was detected on a T cell in the CD4⁺CD32^{high} population, the staining was punctate and often co-localized with punctate CD19 staining (Fig. 2b), suggesting that CD32 was acquired via contact between B and T cells. We noted that the frequency of T cells with punctate CD32 staining was substantially higher in the sorted CD32^{int} population. Thus, sorting for CD4⁺ T cells with a ‘high’ surface density of CD32 results in the selective enrichment of contaminating T–B cell doublets. As shown in Supplementary Fig. 1, these doublets cannot be discerned by routine cytometric FSC and SSC singlet gating strategies.

In summary, using samples from 20 HIV-1-infected, ART-suppressed participants, our data contradict the assertion that CD32a is a marker of

the replication-competent viral reservoir. Although we did detect similar frequencies of CD4⁺CD32^{high} populations to Descours et al.¹, we found no difference in the total HIV-1 DNA content between CD4⁺ T cell populations including or excluding the CD32^{high} fractions (Fig. 1b).

Notably, the CD4⁺CD32^{high} population was highly activated. Previous studies that have evaluated CD32 expression on T cells suggest that it may be detected after activation^{6,7} and led us to believe that this population may be atypical compared to a quiescent population harbouring the HIV-1 reservoir^{2,3}.

Our additional findings are incongruent with CD32a marking the replication-competent reservoir in CD4⁺ T cells; our phenotyping and RT-PCR experiments indicate that it is the CD32b isoform that marks the CD4⁺CD32^{high} cells (Fig. 1d, e, Extended Data Fig. 2b–d). This finding, combined with the demonstration that this cell population is found in uninfected individuals, conflicts with the assertion of Descours et al.¹ that CD32a is upregulated after the establishment of viral latency. Recent reports have corroborated the absence of CD32a transcripts in reactivated, clonal HIV-1-infected CD4⁺ T cells⁸.

The surface density of CD32b (and other B cell markers) on the CD4⁺CD32^{high} population was observed at similar densities to that on B cells. These data, combined with the post-hoc analysis, suggests that this population may be largely comprised of doublets. Direct interrogation of the CD4⁺CD32^{high} population via Amnis imaging confirmed that this population consisted largely of contaminating doublets; either co-incident events or cell-to-cell conjugates (Fig. 2a).

We demonstrate that the mechanism by which the CD32b isoform labels the CD4⁺CD32^{high} populations is through the direct interaction of CD4⁺ T and B cells, and possible trogocytotic transfer of B cell antigens to T cells, as observed in the CD4⁺CD32^{int} population (Fig. 2b). This may explain the transfer or membrane painting of antigens such as CD32b, CD40 and HLA-DR, among other markers^{9–11}. Not only have cell-to-cell membrane transfers been shown to occur commonly *in vivo* during viral infections, but such transfers largely occur on activated cells¹². Membrane-bound Fcγ receptors, including CD32b, are known to be extracted from APCs and then transferred to T cells, and serve as a surrogate of recent T cell and APC interactions¹³. Our demonstration of T–B cell conjugates in the CD4⁺CD32^{high} population and high levels of single cells in the CD4⁺CD32^{int} population support this notion (Fig. 2a, b).

Collectively, our findings confirm that selectively sorting for T cells with a high surface density of CD32 results in the enrichment of T–B cell doublet contaminants, which cannot be discerned by routine gating strategies. The true isoform, CD32b, that marks the CD4⁺CD32^{high} population is probably indicative of dynamic CD4⁺ T cell interaction with B cells, rather than a marker of the HIV-1 reservoir^{14,15}.

We thank S. Mordecai for Amnis technical expertise, and acknowledge support from NIAID grants AI091514, AI122942, AI127089 and AI131365 awarded to J.B.W. Support was also provided by the NIAID awarded Martin Delaney Collaboratory ‘BELIEVE’ grant AI126617, co-funded by NIDA, NIMH and NINDS awarded to D.F.N.

Methods

HIV-1⁺ participants were recruited through: The Maple Leaf Medical clinic in Toronto, Canada; The HIV Eradication and Latency (HEAL) cohort of Brigham and Women’s and Massachusetts General Hospital; The Whitmann Walker Clinic in Washington, DC; or the Hospital of the University of Pennsylvania. The study was approved by the University of Toronto, The University of Pennsylvania and George Washington University ethics committees and according to the protocol approved by the Partners Human Research Committee and Institutional Review Board (IRB). Written informed consent was obtained from each participant.

The percentage of CD32⁺ (clone FUN-2) CD4⁺ T cells was measured in samples from study participants. Both CD32⁺ and CD32^{neg} CD4⁺ T cells were sorted and viral DNA was measured using ddPCR. The analysis of cell lineage markers by flow cytometry and RT-PCR was also conducted. Flow cytometry sorts from PBMCs used in HIV-1 DNA analyses were performed on cell subsets and assessed using Amnis imaging flow cytometry.

Data availability. All data and reagents are available from the corresponding author upon request.

Christa E. Osuna¹, So-Yon Lim¹, Jessica L. Kublin¹, Richard Apps², Elsa Chen¹, Talia M. Mota³, Szu-Han Huang³, Yanqin Ren³, Nathaniel D. Bachtel³, Athe M. Tsibris⁴, Margaret E. Ackerman⁵, R. Brad Jones³, Douglas F. Nixon³ & James B. Whitney^{1,6*}

¹Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. ²Center for Human Immunology, National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA. ³Division of Infectious Diseases, Weill Department of Medicine, Weill Cornell Medical College, New York, NY, USA. ⁴Brigham and Women’s Hospital, Boston, Massachusetts Harvard Medical School, Boston, MA, USA. ⁵Thayer School of Engineering, Dartmouth College, Hanover, NH, USA. ⁶Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA. *e-mail: jwhitne2@bidmc.harvard.edu

Received: 11 August 2017; Accepted: 24 May 2018;
Published online 19 September 2018.

1. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
2. Chun, T. W. et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387**, 183–188 (1997).
3. Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
4. Veri, M. C. et al. Monoclonal antibodies capable of discriminating the human inhibitory Fcγ-receptor IIB (CD32B) from the activating Fcγ-receptor IIA (CD32A): biochemical, biological and functional characterization. *Immunology* **121**, 392–404 (2007).
5. Boruchov, A. M. et al. Activating and inhibitory IgG Fc receptors on human DCs mediate opposing functions. *J. Clin. Invest.* **115**, 2914–2923 (2005).
6. Engelhardt, W., Matzke, J. & Schmidt, R. E. Activation-dependent expression of low affinity IgG receptors FcγRII(CD32) and FcγRIII(CD16) in subpopulations of human T lymphocytes. *Immunobiology* **192**, 297–320 (1995).
7. Sandilands, G. P. et al. Differential expression of CD32 isoforms following alloactivation of human T cells. *Immunology* **91**, 204–211 (1997).
8. Cohn, L. B. et al. Clonal CD4⁺ T cells in the HIV-1 latent reservoir display a distinct gene profile upon reactivation. *Nat. Med.* **24**, 604–609 (2018).
9. Cone, R. E., Sprent, J. & Marchalonis, J. J. Antigen-binding specificity of isolated cell-surface immunoglobulin from thymus cells activated to histocompatibility antigens. *Proc. Natl Acad. Sci. USA* **69**, 2556–2560 (1972).
10. Hwang, I. et al. T cells can use either T cell receptor or CD28 receptors to absorb and internalize cell surface molecules derived from antigen-presenting cells. *J. Exp. Med.* **191**, 1137–1148 (2000).
11. Wetzel, S. A., McKeithan, T. W. & Parker, D. C. Peptide-specific intercellular transfer of MHC class II to CD4⁺ T cells directly from the immunological synapse upon cellular dissociation. *J. Immunol.* **174**, 80–89 (2005).
12. Rosenits, K., Keppler, S. J., Vucikujia, S. & Aichele, P. T cells acquire cell surface determinants of APC via *in vivo* trogocytosis during viral infections. *Eur. J. Immunol.* **40**, 3450–3457 (2010).
13. Daubeuf, S. et al. Preferential transfer of certain plasma membrane proteins onto T and B cells by trogocytosis. *PLoS One* **5**, e8716 (2010).
14. Garside, P. et al. Visualization of specific B and T lymphocyte interactions in the lymph node. *Science* **281**, 96–99 (1998).
15. Okada, T. et al. Antigen-engaged B cells undergo chemotaxis toward the T zone and form motile conjugates with helper T cells. *PLoS Biol.* **3**, e150 (2005).

Author contributions D.F.N. and J.B.W. designed the studies. R.B.J., R.A., E.C., Y.R., N.D.B., C.E.O., R.T. and S.Y.L. led the virology assays. S.H.H., D.C., J.L.K., M.A. and C.E.O. led the immunology assays. J.B.W. led the studies and wrote the paper with all co-authors.

Competing interests Declared none.

Additional information

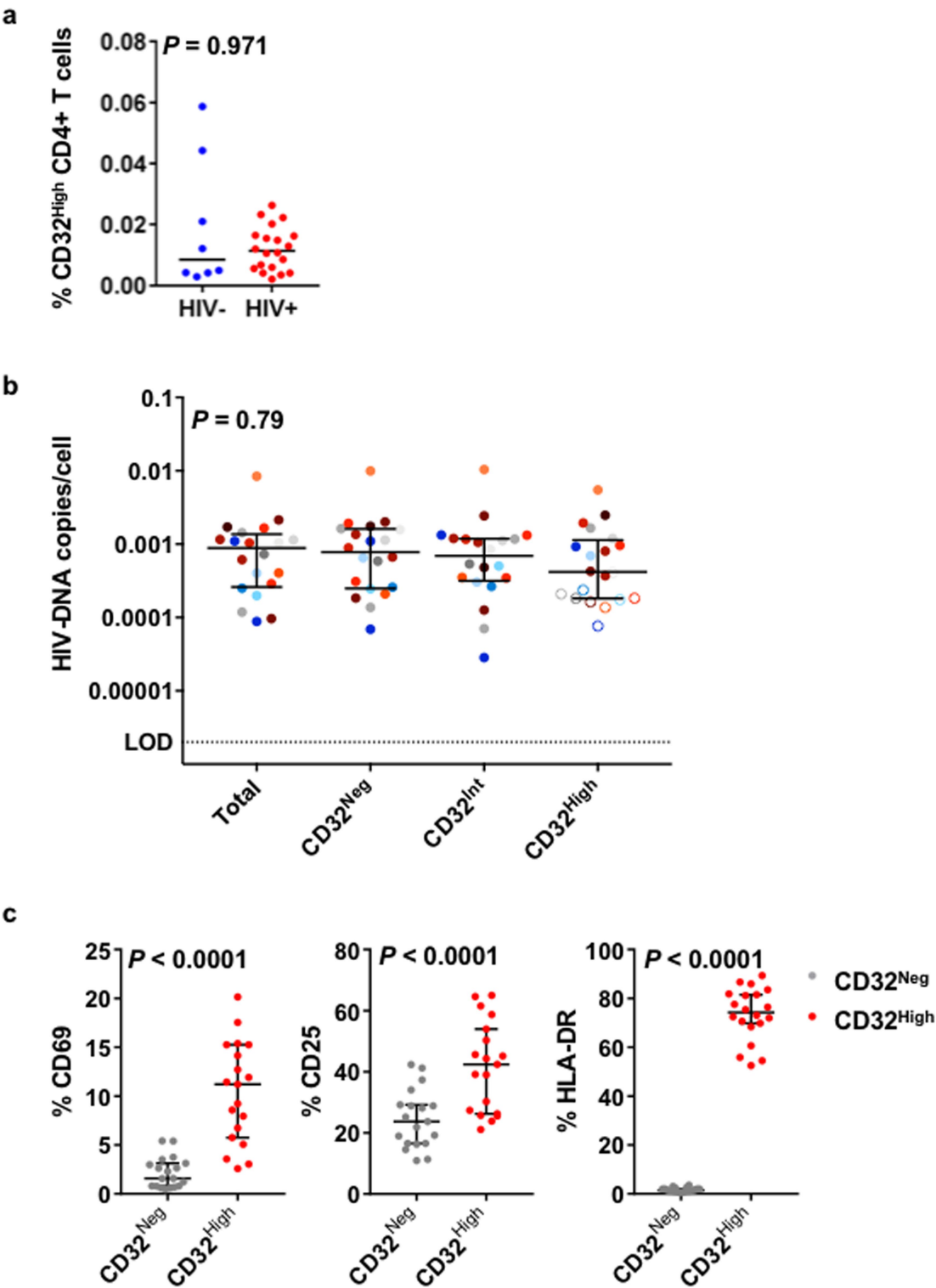
Extended data accompanies this Comment.

Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

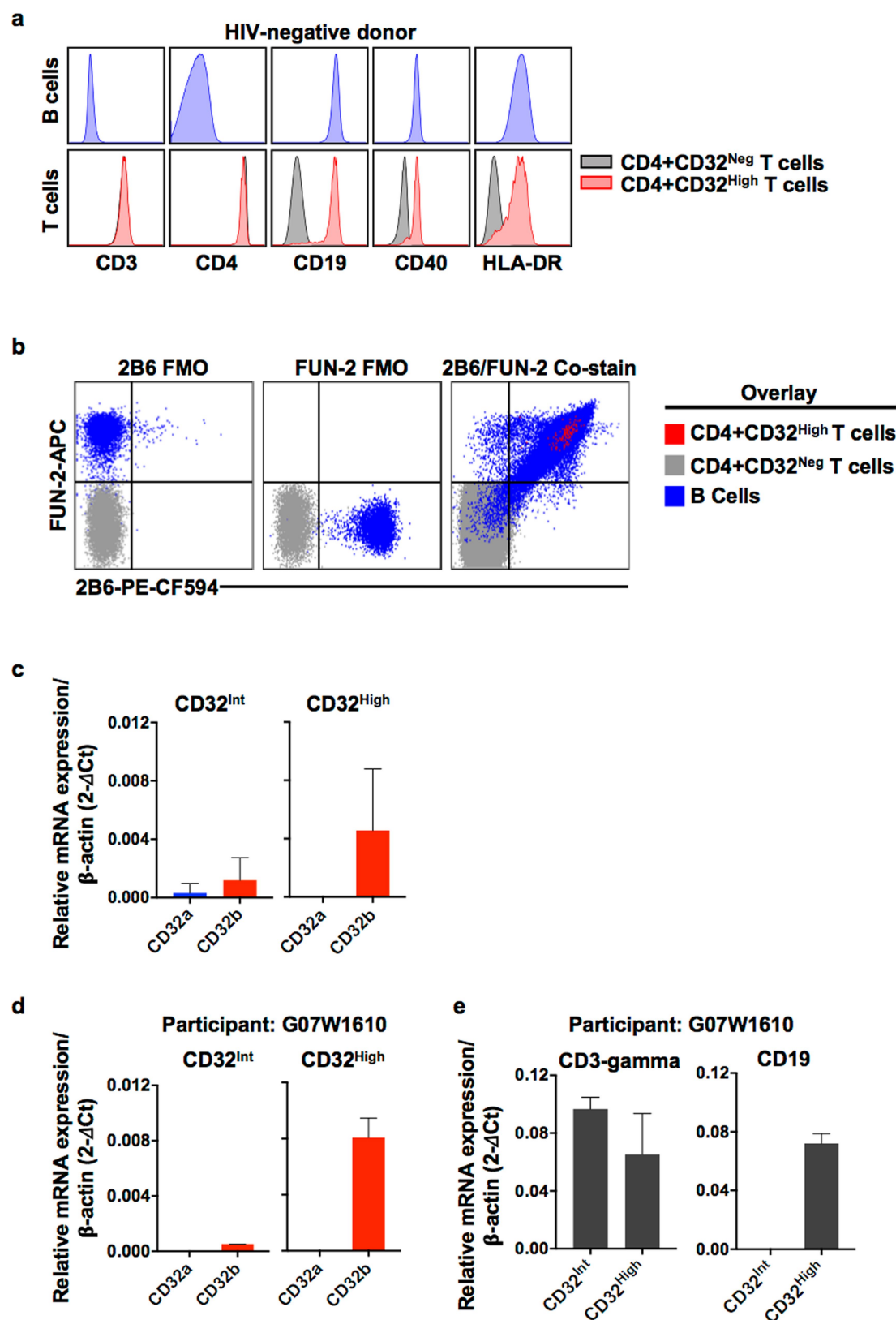
Correspondence and requests for materials should be addressed to J.B.W.

<https://doi.org/10.1038/s41586-018-0495-2>



Extended Data Fig. 1 | Frequency and activation status of CD32-expressing CD4⁺ T cells and their HIV-1 DNA content. **a**, The frequency of CD32^{high} CD4⁺ T cells was measured by flow cytometry in PBMCs from ART-suppressed, HIV-1⁺ ($n = 20$) and HIV-1⁻ ($n = 8$) donors. Bars denote median values. P values were determined by a Mann–Whitney test. **b**, DNA copies per cell in sorted subsets of total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells are shown with median values and the IQR. The results are shown as either the actual HIV-1 DNA

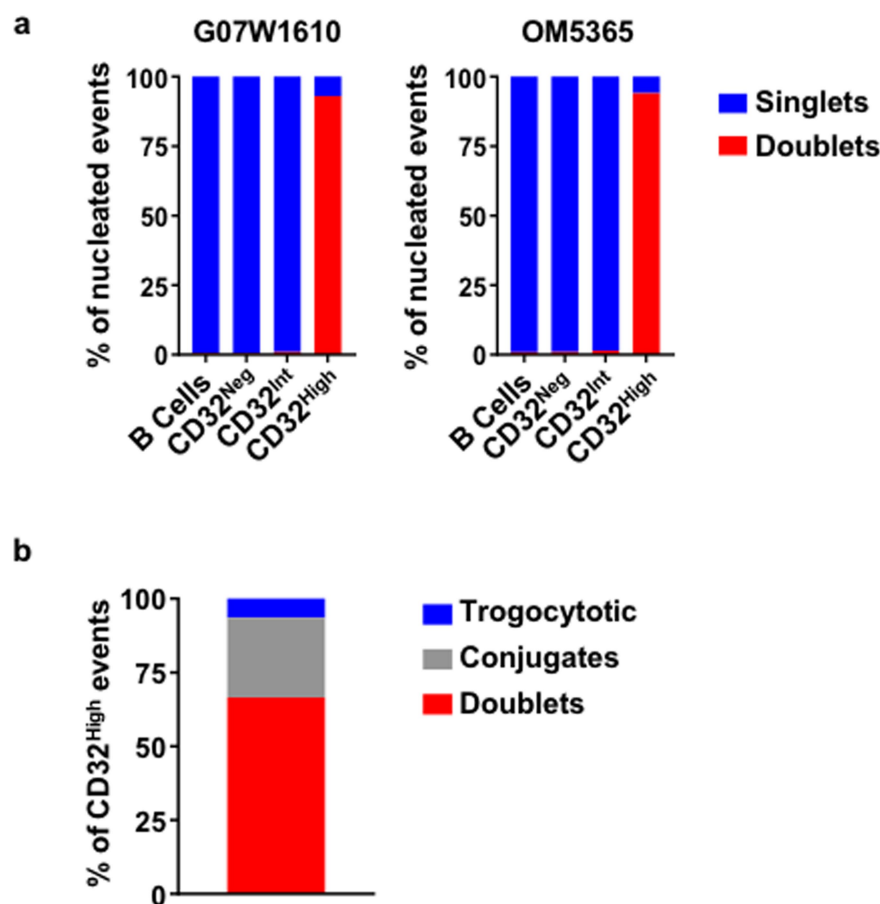
copies per million cells (filled symbols) or as estimated values calculated using the LOD and applied to the number of cells when the DNA input did not reach the threshold (open symbols). P values were determined by a Kruskal–Wallis test. **c**, The percentage of CD69, CD25 and HLA-DR expression was measured by flow cytometry on CD32^{neg} and CD32^{high} (FUN-2) CD4⁺ T cells from PBMCs from HIV-1⁺ participants ($n = 20$). Error bars show the median and IQR. P values were determined by Wilcoxon matched-pairs signed rank tests.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Detection of B cell proteins and mRNA in CD32-expressing CD4⁺ T cells. **a**, CD32^{neg} and CD32^{high} (FUN-2) CD4⁺ T cells from human PBMCs were assessed by flow cytometry for the expression of CD19, CD40 and HLA-DR, and compared to B cells (CD3⁻ CD14⁻ CD19⁺ lymphocytes). Representative flow cytometry results of per cell antigen levels on B cells (top, blue histograms) and CD32^{neg} and CD32^{high} CD4⁺ T cells (bottom, grey and red histograms, respectively) from an HIV-1⁻ donor. **b**, Representative CD32b staining of PBMCs from an HIV-1⁺, ART-suppressed participant. PBMCs were stained with an optimized concentration of the 2B6 monoclonal anti-CD32b antibody, followed by an antibody cocktail that included the FUN-2 monoclonal pan-CD32 antibody, as described in the Methods. Shown are the 2B6 and FUN-2

fluorescence minus one (FMO) antibody cocktail-stained samples and a sample co-stained with 2B6 and FUN-2. **c**, **d**, CD32 mRNA expression levels in CD4⁺CD32⁺ subsets. **c**, The relative expression of CD32A and CD32B mRNA isoforms in sorted CD4⁺CD32^{int} and CD4⁺CD32^{high} subsets from HIV-1⁺, ART-suppressed participants ($n = 4$). **d**, mRNA expression of CD32A and CD32B from patient G07W1610. **e**, T and B cell lineage-specific mRNA transcripts in sorted CD4⁺CD32⁺ subsets from participant G07W1610. Relative mRNA expression of target genes was normalized to *ATCB* using the comparative C_t method. Results are mean \pm s.d. of each value from each participant ($n = 4$; **c**), or from values generated from two separate experiments using samples from the same patient (**d**).



Extended Data Fig. 3 | Doublet composition of the sorted CD4⁺CD32^{high} T cells. Sorted B cells and CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells from an HIV-1⁺, ART-suppressed participant were analysed using an Amnis imaging cytometer. Singlets and doublets were quantified using the aspect ratio and nuclear staining. **a**, The proportion of total singlet and doublet events among total nucleated

cells detected on the Amnis cytometer in each sorted population was determined, and is shown as individual composite bar graphs for two patients (G07W1610 and OM5365). **b**, A composite bar graph of the proportion of conjugates, doublets and trogocytotic events that comprised the sorted CD4⁺CD32^{high} population ($n = 2$).

BRIEF COMMUNICATIONS ARISING

Extended Data Table 1 | Viral suppression of 20 HIV-1-infected participants on ART

Cohort	Participant ID	Date of Initial Suppression (MM/YY)	Length of suppression (yrs)
HEAL	HEAL-009	3/14	3
	HEAL-019	8/09	8
	HEAL-020	3/08	9.3
	HEAL-034	11/05	11.5
	HEAL-053	11/16	1
	HEAL-055	11/00	17
Maple Leaf	CIRC0024	6/98	17.0
	CIRC0133	7/08	7.0
	CIRC0196	4/14	1.2
	OM5011	11/08	6.6
	OM5148	1/08	7.5
	OM5162	9/04	10.8
	OM5203	3/12	3.3
	OM5334	7/14	0.9
	OM5365	3/08	7.3
WWH	WWH-B001	7/11	6.4
	WWH-B005	12/17	0.3
	WWH-B008	11/14	3.1
	WWH-B011	11/11	6
UPenn	G07W1610	10/05	11.8

BRIEF COMMUNICATIONS ARISING

Extended Data Table 2 | CD4⁺CD32^{high} subset proportions and HIV-1 DNA compared to total CD4⁺ and CD32^{neg} CD4⁺ T cells

Participant ID	CD32 ^{High}		HIV-DNA enrichment			
	% in total CD4	Absolute cell count	HIV-DNA copies/cell ¹	CD32 ^{High} /CD4 total ²	CD32 ^{High} /CD32 ^{Neg2}	CD32 ^{Neg} /CD4 total
HEAL-009	0.007	11,427	>0.000002	0.010	0.008	1.236
HEAL-019	0.002	4,911	>0.000002	0.002	0.001	1.500
HEAL-020	0.011	8,482	>0.000002	0.008	0.008	1.036
HEAL-034	0.022	8,238	0.000426	0.199	0.212	0.937
HEAL-053	0.004	9,544	>0.000002	0.017	0.015	1.161
HEAL-055	0.011	21,806	0.00037	0.604	0.555	1.088
CIRC0024	0.015	26,200	>0.000002	0.023	0.029	0.782
CIRC0133	0.017	14,602	>0.000002	0.005	0.010	0.517
CIRC0196	0.006	5,935	0.000694	1.722	1.066	1.615
OM5011	0.008	8,862	0.001942	1.871	2.187	0.855
OM5148	0.004	8,788	0.001191	1.043	1.049	0.994
OM5162	0.016	7,133	0.000923	0.842	0.842	1.000
OM5203	0.026	12,254	>0.000002	0.021	0.011	1.911
OM5334	0.016	6,275	0.000959	0.579	0.499	1.160
OM5365	0.006	11,027	>0.000002	0.003	0.003	0.803
WWH-B001	0.020	10,922	>0.000002	0.007	0.006	1.076
WWH-B005	0.013	5,964	0.00547	0.650	0.550	1.182
WWH-B008	0.023	6,464	0.000805	0.696	0.595	1.169
WWH-B011	0.004	5,953	0.001653	1.154	1.016	1.135
G07W1610	0.012	7,984	0.002482	1.452	1.411	1.029
Median	0.012	8,635	0.000398	0.389	0.356	1.082

¹Values below the LOD (2 copies per 10⁶ cells) are shaded in grey.

²To calculate HIV-1 enrichment, 0.000002 was used for all values below the LOD.

Descours et al. reply

REPLYING TO L. Pérez et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0493-4> (2018); C. E. Osuna et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0495-2> (2018); L. N. Bertagnolli et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0494-3> (2018)

In our previous work¹, we used an in vitro model of HIV-infected unstimulated CD4 T cells to identify CD32 as a candidate marker of HIV⁺ resting CD4 T cells in vitro, and a subset of HIV⁺ total CD4 T cells containing replication-competent viruses in individuals that underwent anti-retroviral therapy (ART). Of note, we did not explore the transcriptional status of hosted viruses (latent or active) ex vivo, nor the activation state of these cells (quiescent or activated)¹. In the accompanying Comments^{2–4}, colleagues attempted to reproduce these findings. They present experiments that support the following conclusions: (1) the isolation of the CD32⁺ CD4 T cell population results from artefacts caused by the flow cytometry sorting method^{2,3}, and (2) the sorted CD32 CD4 T cell population is not enriched in HIV nor in replication-competent proviral DNA^{2–4}. Here, we formulate two questions that mirror the major issues raised by these three Comments^{2–4} and discuss their results in the context of our previous report¹ and more recently published studies.

Is there any evidence that a CD4 T cell can express CD32 in the context of HIV infection? This question is raised by both Osuna et al.² and Pérez et al.³. A recent report⁷, using in situ hybridization (which avoids the criticism of artefacts caused by flow cytometry sorting), showed that HIV-1 RNA co-localized with CD32A (also known as FCGR2A) RNA in 90% of examined cells in B cell follicles from four individuals. Because HIV primarily targets CD4 T cells, these data may support the ability of a CD4 T cell to upregulate CD32 mRNA transcription after infection in vivo. Three independent groups have identified CD32 as being expressed by latently or productively infected CD4 T cells in vitro^{1,5–7}. These models generated and analysed a substantial percentage of HIV-infected CD4 cells. Thus, any marker that is usually not expressed by CD4 T cells but that is detected at the surface of these cells after infection is unlikely to result from biased analyses of cellular doublets, as could be the case when working on rare events from ex vivo samples^{2,3}. Instead, these data suggest that transcriptional regulation leading to the expression of CD32 mRNA and protein can probably occur after in vitro and in vivo infection of a single CD4 T cell.

Does the CD32 CD4 T cell subset contribute to viral persistence under treatment? All three of the accompanying Comments^{2–4} indicate that CD32 CD4 T cells are not enriched for HIV DNA in blood. Recent work suggests, however, that in some virally suppressed HIV-infected individuals, CD32 CD4 T cells were enriched in HIV DNA, although to a lesser extent than we reported⁸. Notably, this question has been recently addressed in tissues, and results seem to be less contrasted than in blood^{7,9,10}. More importantly, they revealed functional properties of these reservoir cells that have not been previously explored^{7,9,10}. As discussed above, a recent report⁷ found that within the B cell follicles of virally suppressed HIV-infected individuals, most of the cells containing HIV RNA and persisting despite treatment were found to express CD32A RNA⁷. This result seems to be in line with other data¹⁰ that indicate that T follicular helper cells, primarily found in these territories, were enriched for HIV DNA and RNA when expressing CD32¹⁰, although at a lower extent than our previous findings¹. In non-lymphoid rectal tissue, CD4 T cells expressing CD32 were also enriched

for both HIV DNA and RNA⁹. Notably, the co-expression of CD32 and HIV RNA reported in these two publications^{9,10} suggests that CD32 marks transcriptionally active infected cells rather than latent cells. Together, these reports support the ability of CD32 to identify a subset of persistent HIV-infected CD4 T cells and suggest that they could contribute to viral persistence under ART in vivo.

In conclusion, we believe that rather than completely ruling out the relevance of CD32 for the identification of a subset of infected cells in vivo and their contribution to HIV persistence, the whole literature, including the three accompanying Comments^{2–4}, opens new technical challenges and questions that we should solve in the near future.

Benjamin Descours, Gael Petitjean and Monsef Benkirane are solely responsible for this Reply. The contributions of the remaining authors from the original Letter¹ were limited to recruiting patients or performing analysis on blinded samples, and thus only Descours, Petitjean and Benkirane have authored this Reply.

Benjamin Descours¹, Gael Petitjean¹ & Monsef Benkirane^{1*}

¹Institut de Génétique Humaine, Laboratoire de Virologie Moléculaire, UMR9002, CNRS, Université de Montpellier, Montpellier, France.

*e-mail: monsef.benkirane@igh.cnrs.fr

1. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
2. Osuna, C. E. et al. Evidence that CD32a does not mark the HIV-1 latent reservoir. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0495-2> (2018).
3. Pérez, L. et al. Conflicting evidence for HIV enrichment in CD32⁺ CD4 T cells. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0493-4> (2018).
4. Bertagnolli, L. N. The role of CD32 during HIV-1 infection. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0494-3> (2018).
5. Iglesias-Ussel, M., Vandergeeten, C., Marchionni, L., Chomont, N. & Romero, F. High levels of CD2 expression identify HIV-1 latently infected resting memory CD4⁺ T cells in virally suppressed subjects. *J. Virol.* **87**, 9148–9158 (2013).
6. Grau-Expósito, J. et al. A Novel single-cell FISH-flow assay identifies effector memory CD4⁺ T cells as a major niche for HIV-1 transcription in HIV-infected patients. *MBio* **8**, e00876-17 (2017).
7. Abdel-Mohsen, M. et al. CD32 is expressed on cells with transcriptionally active HIV but does not enrich for HIV DNA in resting T cells. *Sci. Transl. Med.* **10**, eaar6759 (2018).
8. Martin, G. E. et al. CD32-expressing CD4 T cells are phenotypically diverse and can contain proviral HIV DNA. *Front. Immunol.* **9**, 928 (2018).
9. Hogan, L. E. et al. Increased HIV- transcriptional activity and infectious burden in peripheral blood and gut-associated CD4⁺ T cells expressing CD30. *PLoS Pathog.* **4**, e006856 (2018).
10. Noto, A., Procopio, F., Corpataux, J. M. & Pantaleo, G. CD32⁺PD1⁺ Tfh cells are the major HIV reservoir in long-term art-treated individuals. *J. Virol.* <https://doi.org/10.1128/JVI.00901-18> (2018).

Author contributions B.D., G.P. and M.B. wrote the manuscript.

Competing interests Declared none.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.B.

<https://doi.org/10.1038/s41586-018-0496-1>

Conflicting evidence for HIV enrichment in CD32⁺ CD4 T cells

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

Descours and colleagues¹ reported a marked enrichment for HIV among CD32a⁺ CD4 T cells in people receiving anti-retroviral therapy (ART). This tiny CD32a⁺ population (0.012% of all blood CD4 T cells) contained a median of 0.56 HIV DNA genomes per cell, and accounted for 26.8–86.3% of HIV DNA in CD4 T cells, thus suggesting that targeting CD32a⁺ CD4 T cells might help to clear HIV reservoirs in vivo. Here, we report our unsuccessful attempts to confirm these findings. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0496-1> (2018).

We first used fluorescence-activated cell sorting (FACS) to sort CD4 T cells with high, intermediate and low levels of CD32 staining (CD32^{hi}, CD32^{int} and CD32^{lo}, respectively) from 10 individuals with chronic HIV infection who were receiving ART (mean duration, 8.8 years; range, 2.7–15). We used cell-staining reagents and gating techniques that matched those used by Descours et al.¹ (see Supplementary Methods and Extended Data Fig. 1). As shown in Fig. 1a, we detected no enrichment for HIV DNA in the CD32^{hi} or CD32^{int} CD4 T cells. Moreover, the CD32^{hi} and CD32^{int} subsets combined accounted for no more than 3% of all HIV DNA copies within circulating CD4 T cells in any of the 10 study participants (Fig. 1b). Post-sort flow cytometry of CD32^{hi} and CD32^{int} populations showed heterogeneous patterns that suggested the formation of T cell–B cell or T cell–monocyte conjugates as the origin of most CD32^{hi} or CD32^{int} CD4 T cells, with separation of these conjugates during sorting (Extended Data Fig. 2).

To rule out the possibility that we had inadvertently obtained false negative results either by excluding HIV-infected, CD32⁺ CD4 T cells using tight light scatter gates or by failing to exclude non-T-cell contaminants, we performed parallel sorts on the same 10 samples using an alternative gating scheme. We used a more inclusive light scatter gate as well as markers for B cells, monocytes, dendritic cells and natural killer cells (Extended Data Fig. 3). Events that were CD3⁺ were separated into fractions that were positive for B cell markers (T–B), positive for one or more other non-CD4-T-cell markers (T–other), or negative for all of these, positive for CD4, and CD32^{hi}, CD32^{int} or CD32^{lo}. Neither CD32^{hi} nor CD32^{int} CD4 T cells were enriched for HIV DNA (Fig. 2a). Similarly, we detected no enrichment for HIV DNA in the T–B and

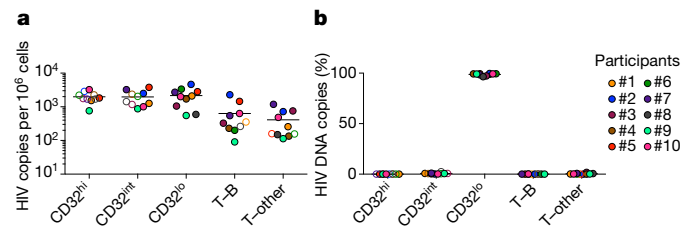


Fig. 2 | Levels of HIV DNA in CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cells, sorted using alternative gating. The samples from Fig. 1 were sorted using alternative gating in which T cells bearing markers of B cells (T–B) or other non-CD4-T-cell lineages (T–other) were first collected in separate tubes. **a**, Copies of HIV DNA per million sorted cells. **b**, Percentages of all HIV DNA copies detected in blood cells that were detected within each subset, calculated by adjusting values in **a** for the relative proportions of these subsets in FACS data.

T–other populations (Fig. 2a). In each of the 10 participants, at least 96% of all HIV DNA copies occurred in conventional CD32^{lo} cells (Fig. 2b). Post-sort flow cytometry suggested that most events bearing both T-cell and non-CD4-T-cell markers again represented cell–cell conjugates, and also showed that most remaining CD32^{hi} CD4 T cells did not reproducibly show a high CD32 signal after sorting (Extended Data Fig. 4). This was in contrast to conventional CD32^{lo} cells, which were uniformly pure in post-sort analyses across participants. In a second group of four individuals whose peripheral blood mononuclear cells (PBMCs) were sorted without previous cryopreservation (Extended Data Fig. 5a), we again found no enrichment for HIV DNA based on CD32 expression (Extended Data Fig. 5b), and also observed that HIV DNA sequences in CD32⁺ CD4 T cells were genetically intermingled with HIV DNA sequences in other CD4 T cells (Extended Data Fig. 5c).

Overall, our studies showed no enrichment for HIV DNA in CD32⁺ CD4 T cells, and also raised questions about the source of the CD32 labelling on these cells. We propose that the CD32 expression associated previously with CD4 T cells could have arisen from adherent non-T-cells or cellular material bearing this marker, and that conjugates containing HIV-infected CD4 T cells could be differentially produced and/or recovered in different laboratories with different sample processing and FACS practices. It is important to acknowledge that these considerations do not explain the discrepancy between the Descours et al. study¹ and ours in the quantities of HIV DNA detected within CD3⁺CD4⁺CD32⁺ sorted material. Nevertheless, we wish to emphasize that our findings do not support targeting CD32 molecules on CD4 T cells in emerging HIV cure strategies.

Methods

Participant recruitment and informed consent were performed under Institutional Review Board (IRB)-approved protocols at the US National Institutes of Health (NIH). For FACS, whole PBMCs were stained with monoclonal antibodies matching those used by Descours et al.¹ (see Supplementary Methods) and sorted on a BD FACSARIA. To evaluate purity, a portion of each population was re-analysed on the flow cytometer after sorting. Virus DNA copies in sorted cells were enumerated by fluorescence-assisted clonal amplification². DNA recovery was quantified by albumin (*ALB*) quantitative PCR. Because the FUN-2 monoclonal antibody used

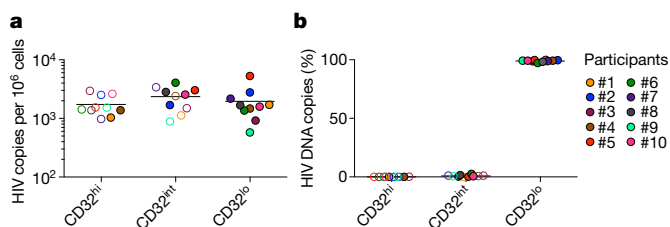


Fig. 1 | Levels of HIV DNA in CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cells, sorted from PBMCs of 10 ART-treated participants, as in Descours et al.¹ **a**, Copies of HIV DNA per million sorted cells. **b**, Percentages of all HIV DNA copies detected in blood CD4 T cells that were detected within each subset, calculated by adjusting values in **a** for the relative proportions of these subsets in FACS data. In all figures, horizontal bars denote median values, and open symbols indicate detection limits for measurements in which HIV DNA was not detected.

BRIEF COMMUNICATIONS ARISING

by Descours et al.¹ and in our study may recognize both CD32a and CD32b, we refer to cells staining with this monoclonal antibody as CD32⁺.

Data availability. All DNA sequences in this manuscript (analysed in Extended Data Fig. 5) have been deposited in GenBank under accession numbers MH080310–MH080572.

Liliana Pérez¹, Jodi Anderson², Jeffrey Chipman³, Ann Thorkelson², Tae-Wook Chun⁴, Susan Moir⁴, Ashley T. Haase⁵, Daniel C. Douek⁶, Timothy W. Schacker^{2,7} & Eli A. Boritz^{1,7*}

¹Virus Persistence and Dynamics Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. ²Division of Infectious Diseases, University of Minnesota, Minneapolis, MN, USA. ³Department of Surgery, University of Minnesota, Minneapolis, MN, USA. ⁴Laboratory of Immunoregulation, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD, USA. ⁵Department of Microbiology and Immunology, University of Minnesota, Minneapolis, MN, USA. ⁶Human Immunology Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. ⁷These authors jointly supervised this work: Timothy W. Schacker, Eli A. Boritz. *e-mail: boritze@mail.nih.gov

Received: 11 October 2017; Accepted: 20 March 2018;

Published online 19 September 2018.

1. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
2. Boritz, E. A. et al. Multiple origins of virus persistence during natural control of HIV infection. *Cell* **166**, 1004–1015 (2016).

Author contributions Data generation and analysis: L.P., J.A., T.W.S. and E.A.B. Study design and oversight: L.P., A.T.H., D.C.D., T.W.S. and E.A.B. Participant cohort and sample management: J.A., J.C., A.T., T.W.C., S.M. and T.W.S. Manuscript preparation: L.P., A.T.H., D.C.D., T.W.S. and E.A.B.

Competing interests Declared none.

Additional information

Extended data accompanies this Comment.

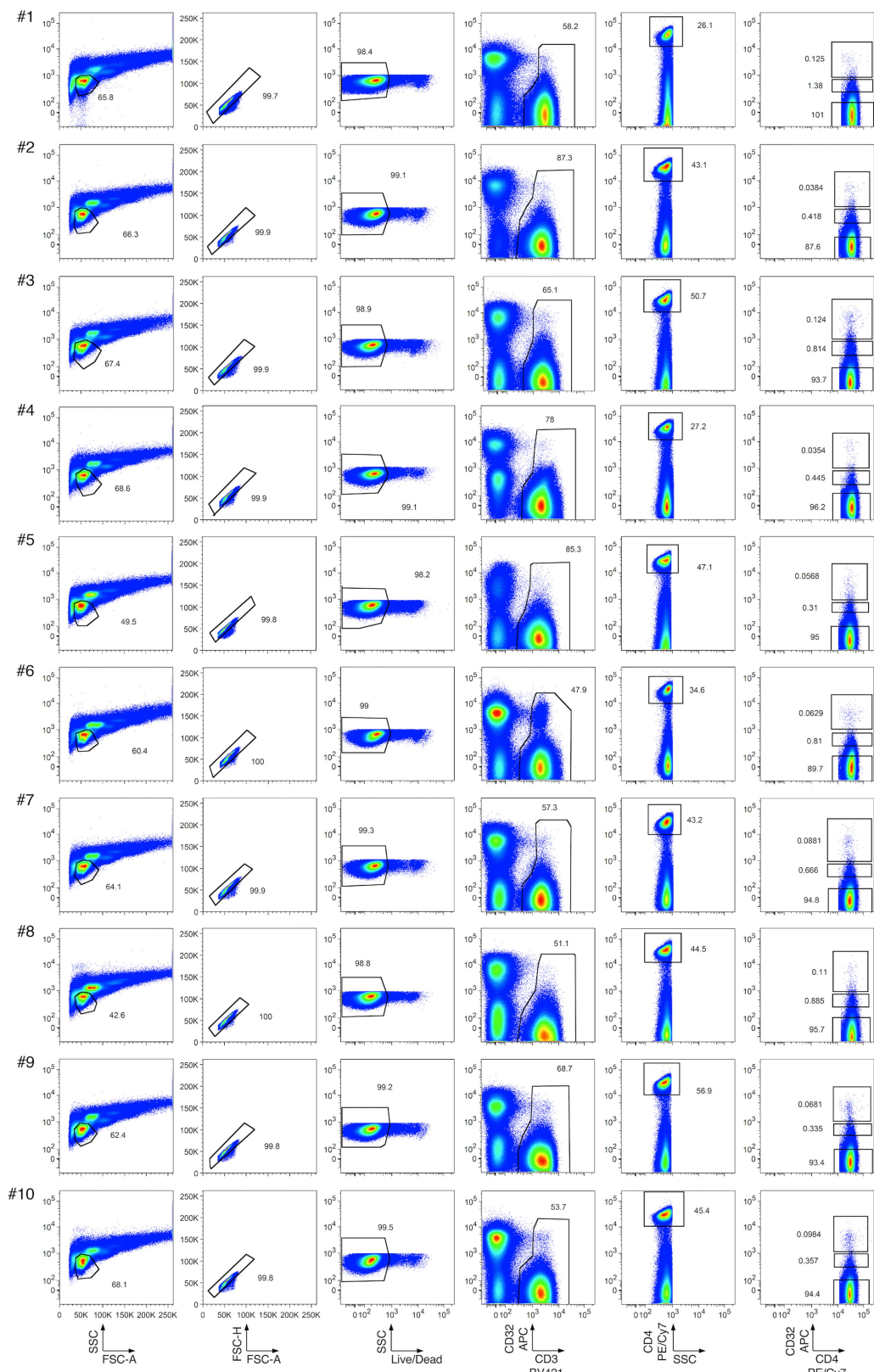
Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

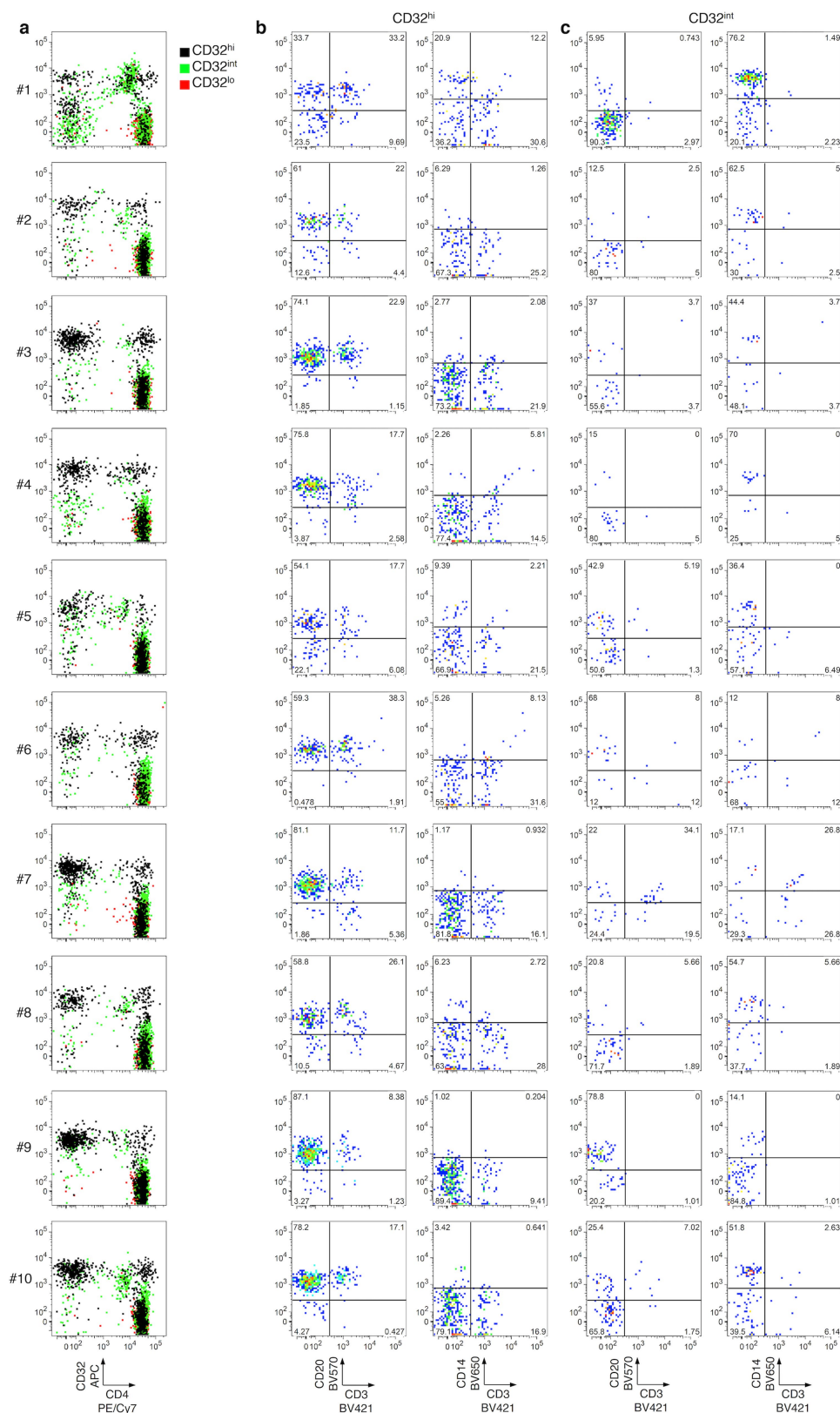
Correspondence and requests for materials should be addressed to E.A.B.

<https://doi.org/10.1038/s41586-018-0493-4>

BRIEF COMMUNICATIONS ARISING



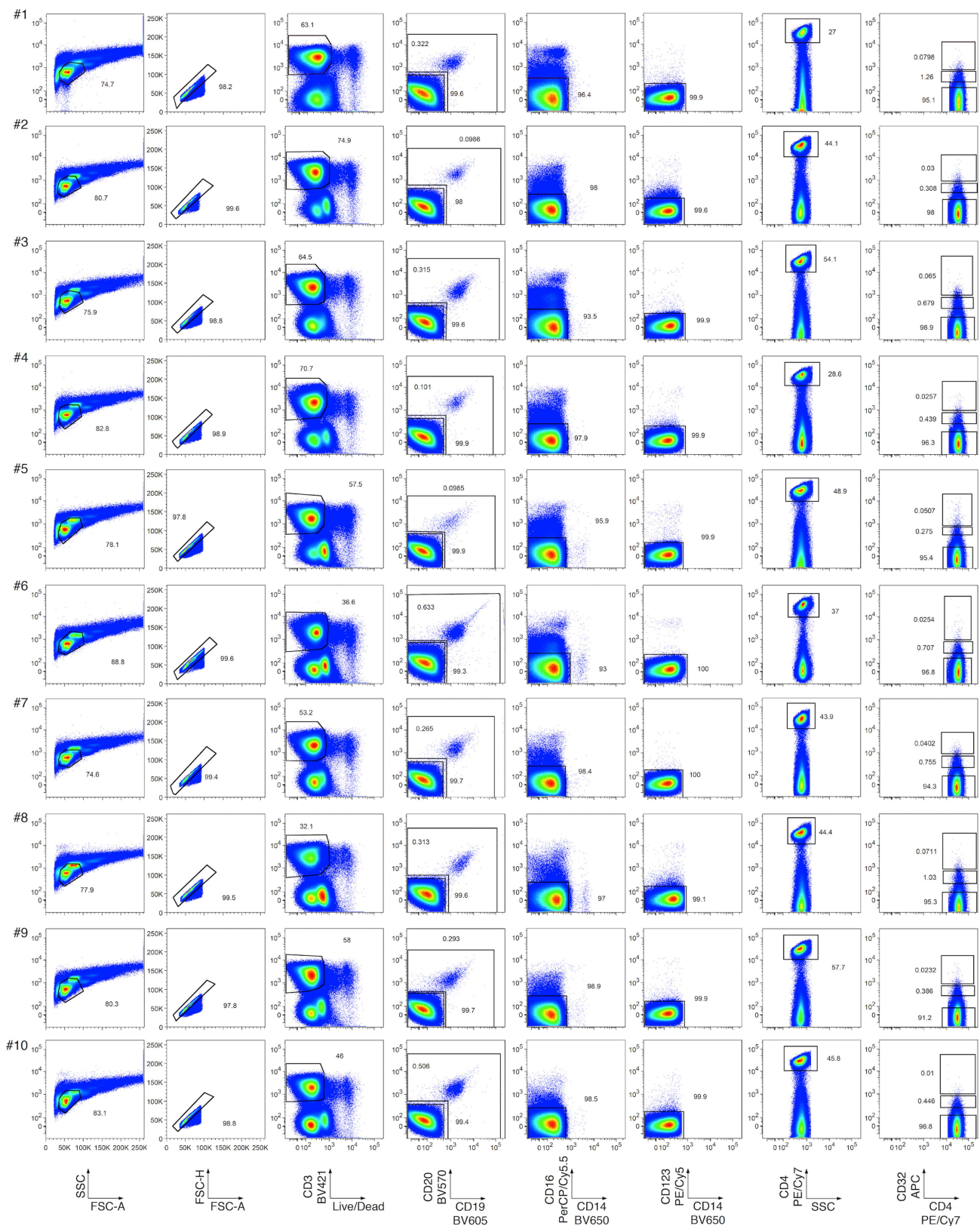
Extended Data Fig. 1 | Flow cytometry of CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations from PBMCs. Single lymphocytes (first two columns) that were viable (third column), CD3⁺ (fourth column), CD4⁺ (fifth column), and CD32^{hi}, CD32^{int} or CD32^{lo} (sixth column) were sorted as described in Descours et al.¹.



Extended Data Fig. 2 | Post-sort flow cytometry of CD32⁺CD4⁺ subsets that were CD32^{hi}, CD32^{int} or CD32^{lo}. Cells were sorted as in Extended Data Fig. 1. **a**, Overlay plots of CD32 and CD4 expression by cells in CD32^{hi}, CD32^{int} and CD32^{lo} sorted populations. Note the heterogeneous pattern of cells from the CD32^{hi} and CD32^{int} populations. **b**, **c**, CD20,

CD14 and CD3 staining in the CD32⁺ cells from the CD32^{hi} (**b**) and the CD32^{int} (**c**) subsets. Note the large proportions of all CD32⁺ cells bearing surface markers consistent with B cells (CD20⁺CD3⁻) or monocytes (CD14⁺CD3⁻) after sorting.

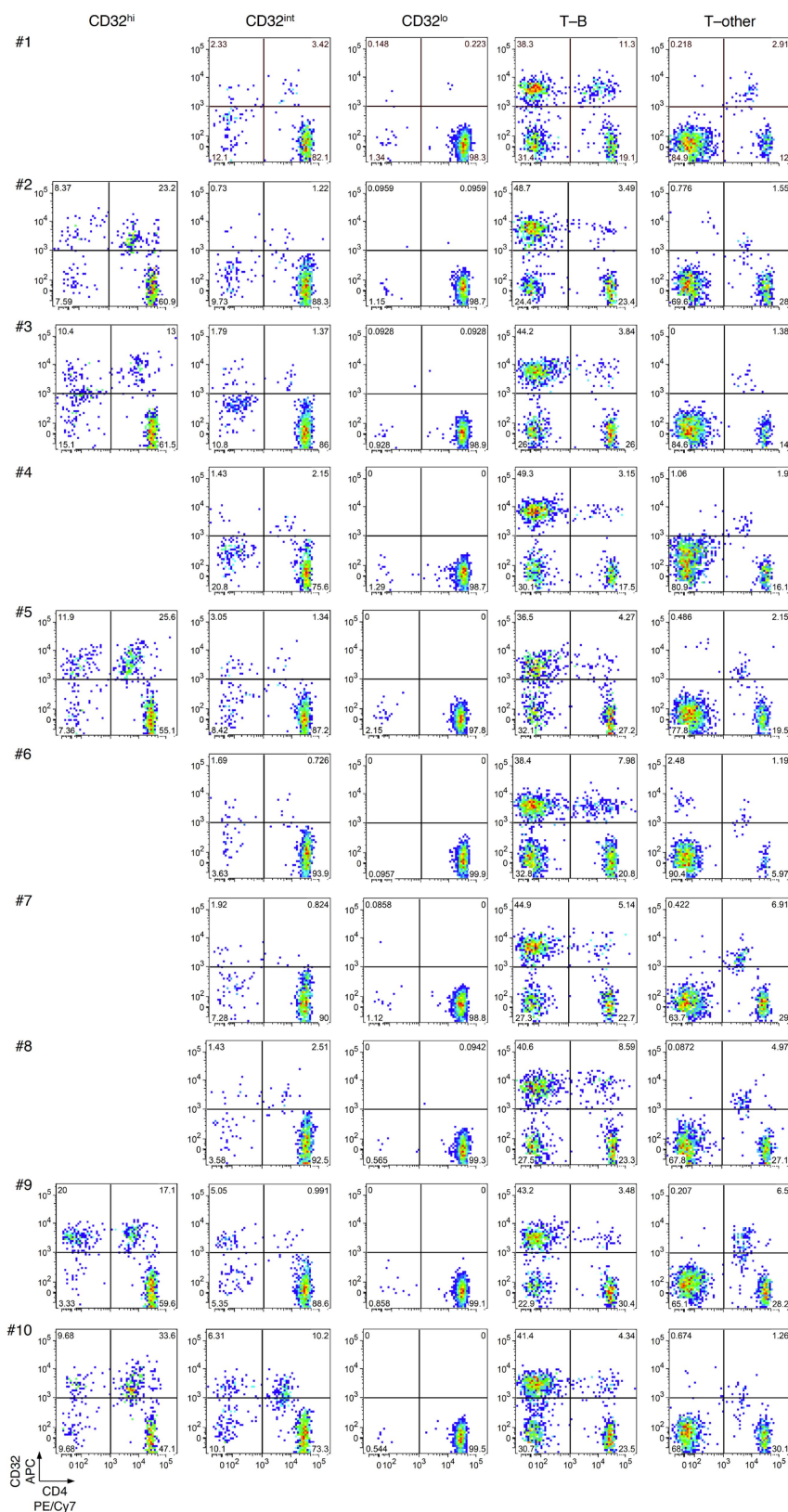
BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 3 | Flow cytometry of PBMCs sorted by alternative gating for CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations, as well as T cell populations bearing markers of B cells (T-B) or other non-CD4-T-cells (T-other). Cells in an inclusive light scatter gate consistent with either small lymphocytes or larger cells (first column) were enriched for single cells (second column). Within these gates, viable CD3⁺ cells

(third column) that were CD19⁻ and CD20⁻ (lower gate, fourth column), CD16⁻ and CD14⁻ (fifth column), CD123⁻ (sixth column), CD4⁺ (seventh column), and CD32^{hi}, CD32^{int} or CD32^{lo} were then collected. Cells that were CD3⁺ and bearing markers of B cells (T-B; upper gate, fourth column) or other non-CD4-T-cells (T-other; combined ungated events from fifth and sixth columns) were also collected in separate tubes.

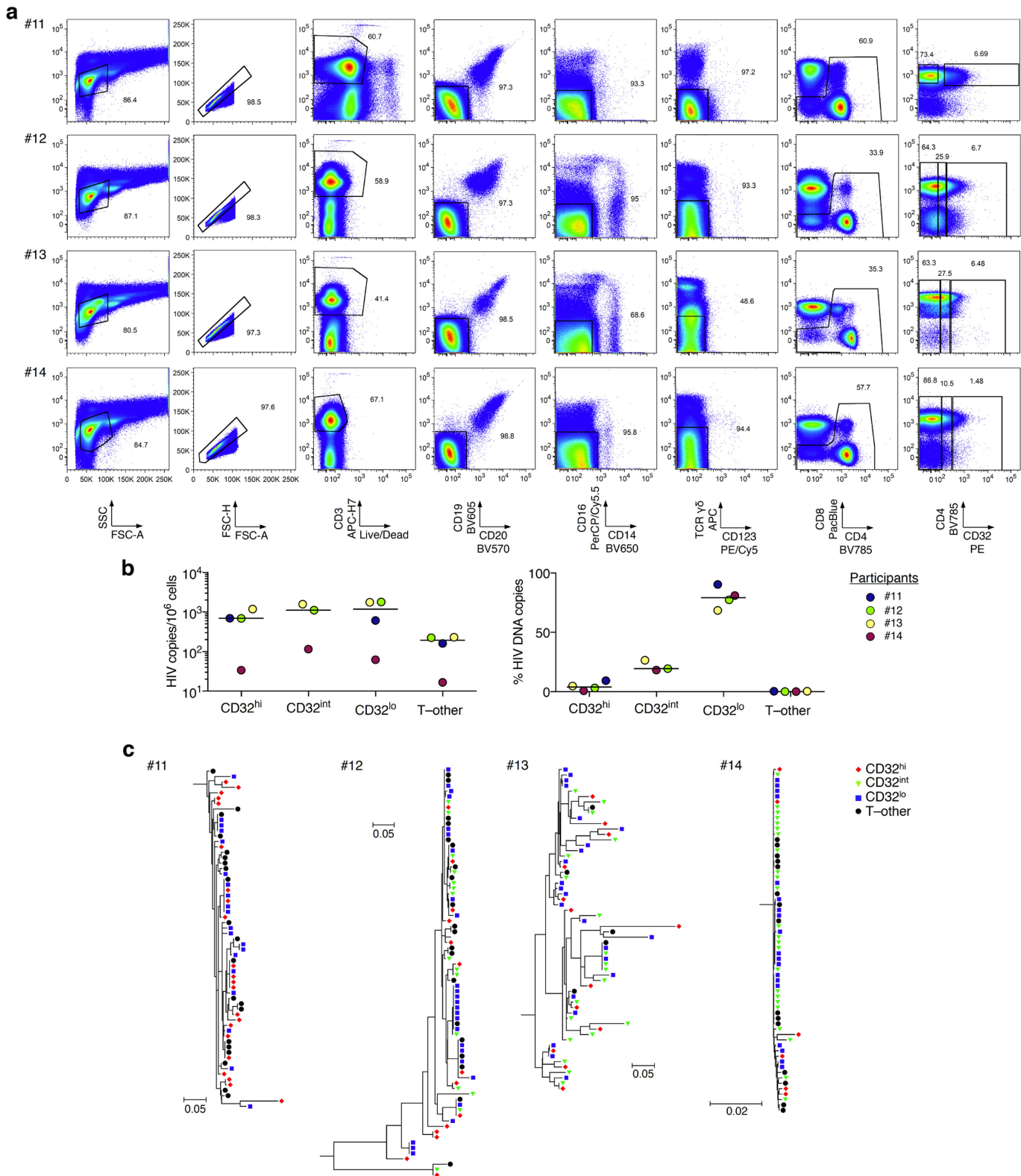
BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 4 | Post-sort flow cytometry of CD32 and CD4 expression by CD32^{hi}, CD32^{int}, CD32^{lo}, T-B and T-other cell subsets. Cells were sorted as in Extended Data Fig. 3. Note the large proportions of all CD32⁺ cells that did not show high CD4 expression after sorting.

Post-sort analyses of CD3⁺CD4⁺CD32^{hi} populations were deferred in cases in which these populations were too small to permit both post-sort analysis and downstream HIV DNA quantification (that is, donors # 1, 4 and 6–8).

BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 5 | See next page for caption.

BRIEF COMMUNICATIONS ARISING

Extended Data Fig. 5 | Flow cytometry, HIV DNA levels, and single-copy HIV DNA sequence analysis from CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations, and from T cells also bearing non-CD4-T-cell markers. **a**, PBMCs from four additional study participants were collected from whole blood by venipuncture with immediate processing (without cryopreservation). The T-other population was collected as a combination of the ungated events from CD19/CD20, CD16/CD14 and $\gamma\delta$ T cell receptor/CD123 exclusion plots (fourth, fifth and sixth columns). **b**, Left, copies of HIV DNA per million cells sorted from four additional study participants as in **a**. Right, percentages of all HIV DNA copies detected in

blood cells deriving from CD32^{hi}, CD32^{int}, CD32^{lo} and T-other subsets, calculated by adjusting values in the left panel for the relative proportions of these subsets determined using FACS data. **c**, Sequences of individual HIV DNA copies were determined by Sanger sequencing of products obtained by fluorescence-assisted clonal amplification, which amplifies a region of the HIV *env* gene. Phylogenetic trees were constructed as described in the Supplementary Methods. All Bonferroni-corrected Slatkin–Maddison *P* values for genetic compartmentalization between any two subsets were greater than 0.05 in all four participants.

The role of CD32 during HIV-1 infection

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

The persistence of latent HIV-1 in resting memory CD4⁺ T cells is a major barrier to a cure, and a biomarker for latently infected cells would be of great scientific and clinical importance^{1–5}. Using an elegant discovery-based approach, Descours et al.⁶ reported that CD32a, an Fcγ receptor not normally expressed on T cells, is a potential biomarker for the HIV-1 reservoir in CD4⁺ T cells⁶. Using a quantitative viral outgrowth assay (qVOA), we show that CD32⁺CD4⁺ T cells do not contain the majority of intact proviruses in the latent reservoir and that the enrichment found by Descours et al.⁶ may in part reflect the use of an ultrasensitive ELISA that does not predict exponential viral outgrowth. Our studies show that CD32 is not a biomarker for the major population of latently infected CD4⁺ T cells. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0496-1> (2018).

If CD32a is a biomarker for latent HIV-1 infection in CD4⁺ T cells, one that is never expressed on CD4⁺ T cells in the absence of HIV-1 infection, then a difference in the frequency of CD4⁺ T cells that express CD32 in HIV-1-infected individuals relative to the frequency in healthy donors is expected. We isolated CD4⁺ T cells from infected and uninfected donors by negative selection and analysed the expression of CD32 and CD4 by flow cytometry. In healthy donors, an average of 0.019% of CD4⁺ T cells was also CD32⁺ (Fig. 1a). This value is not significantly different from levels in HIV-1-infected individuals (Fig. 1a; average 0.011%, $P = 0.1143$) or from values previously reported by Descours et al.⁶ in HIV-1-infected individuals (0.016%, $P = 0.66$). Thus, CD32 does not seem to be a specific biomarker of latently infected CD4⁺ T cells.

To examine whether replication-competent proviruses were present in CD4⁺CD32^{hi} T cells, total CD4⁺ T cells were isolated by negative selection from six HIV-1⁺ individuals that were treated with suppressive anti-retroviral therapy (ART) for at least 6 months (Supplementary Table 1). Freshly isolated cells were stained and sorted to obtain CD4⁺CD32^{hi} and CD4⁺CD32[−] populations, which were analysed in qVOAs⁷ (Fig. 1b, protocol 1). The number of CD4⁺CD32^{hi} cells assayed for each subject is shown in Fig. 1c. On day 14, outgrowth was measured using a standard ELISA for the HIV-1 p24 antigen. CD4⁺CD32^{hi} wells from all subjects were negative for p24 on day 14, and remained negative after an additional week of culture. Conversely, outgrowth was observed in CD4⁺CD32[−] wells from all subjects on both days 14 and 21. The mean infected cell frequency, 1.37 infectious units per million cells (IUPM), was comparable to values previously measured in resting CD4⁺ T cells in several studies (0.03–3.00 IUPM in HIV-1-infected patients⁸, 0.97 IUPM in chronically infected patients⁹) and to values previously measured in the same subjects (mean value 1.33 IUPM) (Fig. 1d, Supplementary Table 2). If the enrichment of proviruses in CD32⁺ cells reported by Descours et al.⁶ was characteristic of replication-competent proviruses, then outgrowth from CD4⁺CD32^{hi} T cells should have been seen (Fig. 1e).

One possible explanation for the discrepancy between our results and those of Descours et al.⁶ is that some latent HIV-1 may be present in a previously undescribed population of CD4⁺ T cells that express CD32 together with other non-T-cell lineage markers. Such cells would be removed during the negative selection used to isolate CD4⁺ T cells. Therefore, we freshly isolated total CD4⁺ cells from infected donors on suppressive ART using two methods: negative selection to remove other lineages, leaving untouched CD4⁺ T cells, and positive selection for

cells expressing CD4 (Fig. 1b, protocol 2). Both CD4⁺ populations were analysed by qVOA. No significant differences were observed in the frequencies of latently infected cells (Fig. 1f). Furthermore, no significant differences in proviral DNA were observed between the purified cell populations (Fig. 1g). Because CD4 is required for HIV-1 entry into the host cell, cell populations obtained via positive selection for CD4 should include every latently infected CD4⁺ T cell. Given that neither the infected cell frequencies nor the levels of proviral DNA differed between the purified cell populations, we conclude that no additional sizable population of latently infected cells was recovered by positive CD4 selection.

In further studies, we used a cell sorting strategy identical to that of Descours et al.⁶ on samples freshly isolated from six subjects receiving ART treatment. Peripheral blood mononuclear cells (PBMCs) isolated from subjects were stained and sorted to obtain CD3⁺CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32[−] cell populations that were tested for latently infected cells by qVOA analysis. The numbers of CD3⁺CD4⁺CD32^{hi} cells assayed for each subject are shown in Fig. 1c and Supplementary Table 3. In addition, total CD4⁺ cells were obtained by staining PBMCs for CD4 and sorting for CD4⁺ cells (Fig. 1b, protocol 3). qVOA results showed that both the CD3⁺CD4⁺CD32[−] and the total CD4⁺ T cell populations had the same infected cell frequencies that were comparable to frequencies measured in other studies¹⁰. However, we observed no outgrowth in CD3⁺CD4⁺CD32^{hi} cultures (Fig. 1h, Supplementary Table 2).

We also analysed CD3⁺CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32[−] cells isolated by the method of Descours et al.⁶ for the presence of proviral DNA by qPCR. We found 89 copies of *gag* per million CD3⁺CD4⁺CD32[−] cells, which is similar to previous measurements in total CD4⁺ T cells¹¹. However, no proviral DNA was detected after DNA extraction from 39,000 CD3⁺CD4⁺CD32^{hi} cells and subsequent qPCR analysis (data not shown). This finding makes it highly unlikely that this cell population is enriched for HIV-1 to a level of more than one provirus copy per cell, as reported by Descours et al.⁶. We caution that the normalization of very low-level HIV-1 DNA measurements from qPCR reactions done with a low number of input cells could artificially produce apparent enrichments in HIV-1 DNA.

In a further attempt to explain the discordant qVOA results obtained in our studies and those of Descours et al.⁶, we tested whether the use of the ultra-sensitive p24 digital ELISA¹² and the low cell input can affect IUPM calculations, leading to erroneous overestimation of latent infection. qVOA culture supernatants were assayed for HIV-1 p24 using the ultrasensitive SIMOA p24 2.0 assay (Quanterix) on days 5, 9, 14 and 21. Using the lower limit of quantification (0.01 pg ml^{−1}) as the cut-off level, we found that two out of three qVOAs containing CD4⁺CD32^{hi} cells tested positive for p24 by this assay, even though the same wells were negative by standard ELISA, which is several orders of magnitude less sensitive (Fig. 2a). Exponential outgrowth is the hallmark of replication-competent viruses. In qVOA cultures of CD4⁺CD32[−] cells, only a fraction of the wells that were positive by SIMOA showed exponential outgrowth as determined by standard ELISA on day 21 (Fig. 2b). Importantly, CD4⁺CD32^{hi} culture wells that tested positive by SIMOA p24 assay showed no exponential outgrowth and had significantly lower levels of p24 (Fig. 2c). It is possible that low positive SIMOA values could reflect an assay artefact or the presence of defective proviruses that are still capable of producing low levels of Gag¹³. A further concern

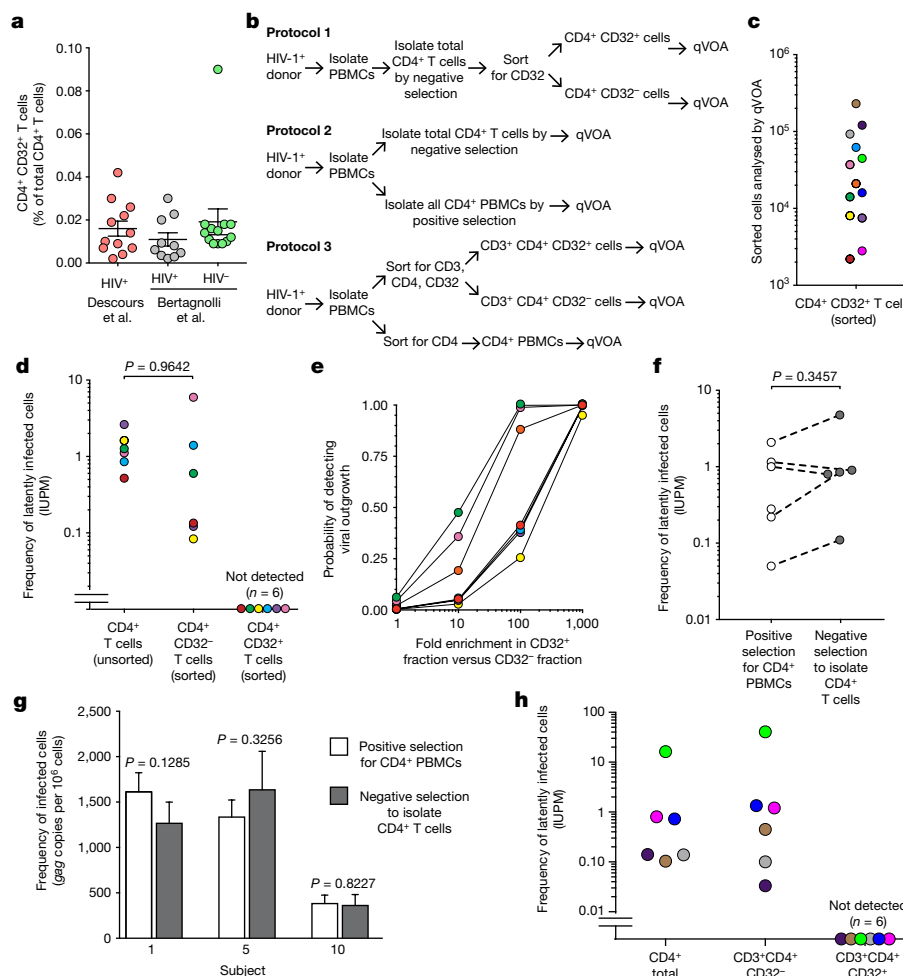


Fig. 1 | Analysis of CD4⁺CD32⁻ and CD4⁺CD32⁺ populations by qVOA and proviral DNA measurements. **a**, Percentage of CD4⁺CD32^{hi} T cells relative to total CD4⁺ T cells in healthy donors and HIV-1-infected donors. Infected donor values were obtained from supplementary table 4 of Descours et al.⁶. LLOQ, lower limit of quantification. **b**, Schematic depicting the three strategies (protocols 1–3) used to obtain different populations of CD4⁺ T cells analysed in qVOAs. **c**, Numbers of sorted CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32^{hi} T cells from each subject analysed in qVOAs. **d**, Frequencies of latently infected cells among CD4⁺CD32^{hi} T cells and CD4⁺CD32⁻ T cells and among total CD4⁺ T cells from the same subjects previously measured in separate experiments. Cells were isolated using protocol 1 (colours correspond to subject values from panel c). **e**, Probability of detecting outgrowth based on measured frequencies of latently infected cells among the CD4⁺CD32⁻ fraction and number of CD4⁺CD32^{hi} cells plated assuming various degrees of enrichment of HIV-1 in CD32^{hi} cells. **f**, Frequencies of latently infected cells measured in qVOAs using positive or negative selection to obtain total CD4⁺ cells (protocol 2; positive selection was accomplished by either sorting or CD4 microbead strategies, with similar results). **g**, Comparison of proviral DNA measurements obtained with qPCR on total CD4⁺ cells purified using positive or negative selection (protocol 2). **h**, Frequencies of latently infected cells among total CD4 cells, and CD3⁺CD4⁺CD32⁻ and CD3⁺CD4⁺CD32^{hi} populations. Cells were isolated using protocol 3 (colours correspond to subject values from panel c).

is that the IUPM calculations are based on cell input, fold dilutions and technical replicates¹⁴, and thus, qVOA analyses performed with very small numbers of sorted CD4⁺CD32^{hi} cells can markedly skew the frequency of cells harbouring replication-competent proviruses (five-fold dilutions from 800 to 1 cell in Descours et al.⁶). When we applied the results obtained with the SIMOA p24 assay, IUPM values ranged from 0 to 3,134 and 554 (patients 4 and 5, respectively; Fig. 2d). As a consequence, when we calculated the ‘fold enrichment’ of IUPM in the CD4⁺CD32^{hi} cells compared to the CD4⁺CD32⁻ cells, we observed a mean fold enrichment of 665 (range 152–1179, from the two patients with positive p24 using SIMOA), similar to what was reported by Descours et al.⁶ (Fig. 2e).

In summary, we find no evidence that CD32 expression indicates the presence of latent HIV-1, and demonstrate that at least a substantial fraction of the HIV-1 latent reservoir is in CD3⁺CD4⁺CD32⁻

T cells. Although no outgrowth could be found in cultures containing CD4⁺CD32^{hi} T cells, viral outgrowth comparable to historical measurements was found in cultures containing CD4⁺CD32⁻ T cells. The use of an ultrasensitive p24 ELISA assay may account for the apparent enrichment observed in culture experiments by Descours et al.⁶. In short, our results have demonstrated that CD32 does not define the HIV-1 reservoir and that future research is needed to identify biomarkers for latently infected cells.

We thank the study participants without whom this research would not be possible. Funding was provided by the US National Institutes of Health (NIH) Martin Delaney I4C, Beat-HIV and DARE Collaboratories by the Johns Hopkins Center for AIDS Research (P30AI094189), by NIH grant 43222, and by the Howard Hughes Medical Institute and the Bill and Melinda Gates Foundation.

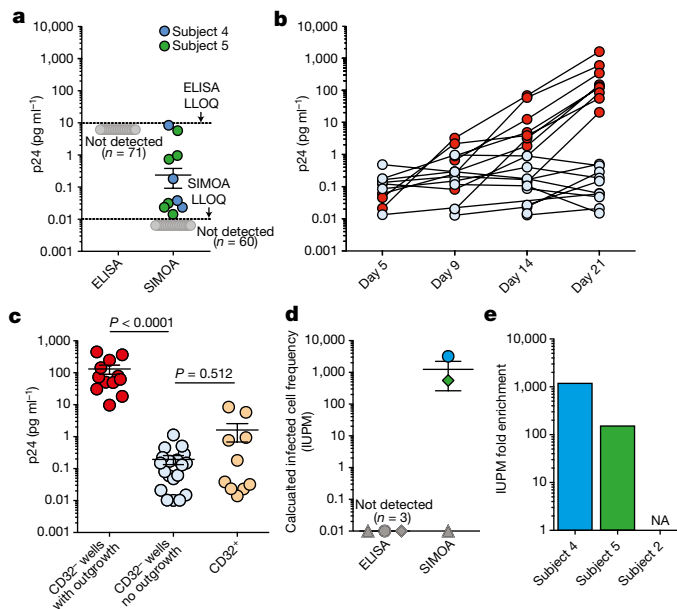


Fig. 2 | Ultrasensitive p24 measurements. **a**, Levels of p24 from CD32⁺ culture wells measured by ELISA and SIMOA (lower limit of quantification: 5–10 pg ml⁻¹ and 0.01 pg ml⁻¹, respectively) (data collected from three subjects, for a total of 71 wells). **b**, Longitudinal levels of p24 measured by SIMOA in individual culture wells in the qVOA for CD32⁻ cells from subject 5, showing wells with and without viral outgrowth (red and blue circles, respectively). **c**, Levels of p24 measured by ELISA in CD32⁻ wells with outgrowth compared with SIMOA measurements in wells with no outgrowth and CD32⁺ wells (data collected from subjects 2, 4 and 5). *P* values were determined with a non-parametric *t*-test. **d**, IUPM calculation based on ELISA and SIMOA analysis. Symbols in dark grey represent values below the limit of detection. **e**, Fold enrichment of IUPM in CD32⁺ cells (from subjects 2, 4 and 5). NA, not applicable.

Methods

qVOAs isolated CD4⁺ T cells using negative depletion and were sorted for CD32⁺ cells (Fig. 1b, protocol 1). To test whether negative depletion was causing a loss of CD32⁺ CD4⁺ T cells, outgrowth and proviral DNA were compared from qVOAs in which CD4⁺ T cells were isolated using positive selection to measurements using negative depletion. Outgrowth measurements and proviral DNA were also measured using the methods described by Descours et al.⁶. Proviral DNA measurements were performed using qPCR¹⁵. HIV-1 p24 values were measured using both a standard ELISA for p24 antigen (Perkin Elmer) and SIMOA (Quanterix). Further details are provided in Supplementary Methods.

Data availability. All data are available from the corresponding author upon reasonable request.

Lynn N. Bertagnoli¹, Jennifer A. White¹, Francesco R. Simonetti¹, Subul A. Beg^{1,2}, Jun Lai^{1,2}, Costin Tomescu³, Alexandra J. Murray¹, Annukka A. R. Antar¹, Hao Zhang⁴, Joseph B. Margolick⁴, Rebecca Hoh⁵, Stephen G. Deeks⁵, Pablo Tebas⁶, Luis J. Montaner³, Robert F. Siliciano^{1,2*}, Gregory M. Laird¹ & Janet D. Siliciano¹

¹Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ²Howard Hughes Medical Institute, Baltimore, MD, USA. ³The Wistar Institute, Philadelphia, PA, USA. ⁴Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. ⁵Division of HIV, Infectious Diseases and Global Medicine, University of California, San Francisco, CA, USA. ⁶Division of Infectious Diseases, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. *e-mail: rsiliciano@jhmi.edu

Received: 29 September 2017; Accepted: 3 April 2018;

Published online 19 September 2018.

1. Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
2. Chun, T. W. et al. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
3. Wong, J. K. et al. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
4. Richman, D. D. et al. The challenge of finding a cure for HIV infection. *Science* **323**, 1304–1307 (2009).
5. Deeks, S. G. et al. Towards an HIV cure: a global scientific strategy. *Nat. Rev. Immunol.* **12**, 607–614 (2012).
6. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
7. Laird, G. M., Rosenbloom, D. I., Lai, J., Siliciano, R. F. & Siliciano, J. D. Measuring the frequency of latent HIV-1 in resting CD4⁺ T cells using a limiting dilution coculture assay. *Methods Mol. Biol.* **1354**, 239–253 (2016).
8. Siliciano, J. D. et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4⁺ T cells. *Nat. Med.* **9**, 727–728 (2003).
9. Eriksson, S. et al. Comparative analysis of measures of viral reservoirs in HIV-1 eradication studies. *PLoS Pathog.* **9**, e1003174 (2013).
10. Crooks, A. M. et al. Precise quantitation of the latent HIV-1 reservoir: implications for eradication strategies. *J. Infect. Dis.* **212**, 1361–1365 (2015).
11. Besson, G. J. et al. HIV-1 DNA decay dynamics in blood during more than a decade of suppressive antiretroviral therapy. *Clin. Infect. Dis.* **59**, 1312–1321 (2014).
12. Passaes, C. P. & Sáez-Cirión, A. HIV cure research: advances and prospects. *Virology* **454–455**, 340–352 (2014).
13. Pollack, R. A. et al. Defective HIV-1 proviruses are expressed and can be recognized by cytotoxic T lymphocytes, which shape the proviral landscape. *Cell Host Microbe* **21**, 494–506.e4 (2017).
14. Rosenbloom, D. I. et al. Designing and interpreting limiting dilution assays: general principles and applications to the latent reservoir for human immunodeficiency virus-1. *Open Forum Infect. Dis.* **2**, ofv123 (2015).
15. Massanella, M., Gianella, S., Lada, S. M., Richman, D. D. & Strain, M. C. Quantification of total and 2-LTR (long terminal repeat) HIV DNA, HIV RNA and herpesvirus DNA in PBMCs. *Bio Protoc.* **5**, e1492 (2015).

Author contributions L.N.B., J.A.W., G.M.L., R.F.S. and J.D.S. designed experiments. S.A.B., C.T. and L.J.M. obtained samples. L.N.B., J.A.W., S.A.B., G.M.L., R.F.S., J.L., A.J.M., A.A.R.A. and J.D.S. performed experiments. R.F.S., H.J. and J.B.M. performed cell sorting. L.N.B., J.A.W., R.F.S., A.J.M., A.A.R.A., R.F.S. and J.D.S. analysed the data and wrote the manuscript.

Competing interests Declared none.

Additional information

Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.F.S.

<https://doi.org/10.1038/s41586-018-0494-3>

Evidence that CD32a does not mark the HIV-1 latent reservoir

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

A recent report by Descours et al.¹ suggests that the cell surface expression of the low affinity Fc receptor CD32a (also known as FcγRIIa) marks the replication-competent HIV-1 reservoir in CD4⁺ T cells from 12 HIV-1-infected participants receiving suppressive anti-retroviral therapy (ART)¹. We have undertaken considerable efforts to replicate these findings using peripheral blood mononuclear cells (PBMCs) from 20 HIV-1-infected, ART-suppressed participants (Extended Data Table 1). We found no evidence to suggest that CD32a marks a CD4⁺ T cell population enriched in either HIV-1 DNA or replication-competent HIV-1 in our study participants. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-s41586-018-0496-1> (2018).

To validate these findings, we adopted the same gating strategy as described by Descours et al.¹ to define CD4⁺ T cell populations (Supplementary Fig. 1a). The CD32 antigen was identified using the same antibody clone (FUN-2) as described by Descours et al.¹. We observed the same CD4⁺ T cell subsets that stained at a high cell surface density of CD32 (CD4⁺CD32^{high}), an intermediate cell surface density of CD32 (CD4⁺CD32^{int}), and a CD4⁺ T cell subset lacking CD32 expression (CD4⁺CD32^{neg}). We obtained frequencies of CD4⁺CD32^{high} T cells that ranged from 0.002% to 0.026%, with a median value (0.012%) that was identical to that reported by Descours et al.¹ (Extended Data Table 2 and Supplementary Fig. 1a). Notably, we confirmed that this same CD4⁺CD32^{high} population is also present in PBMCs isolated from eight healthy donors and exists at similar frequencies to that in HIV-1-infected samples ($P = 0.971$, Extended Data Fig. 1a).

Next, we assessed the amount of replication-competent HIV-1 isolated from the same 20 participants by measuring the infectious unit per million cells (IUPM) in CD4⁺ T cells (range 0.01–37.5, median 0.46). Participant CD4⁺CD32^{high} T cell populations were colour-coded in descending order, and then divided into quartiles that corresponded to the relative frequency of CD4⁺CD32^{high} cells present in these samples (Fig. 1a).

After cytometric sorting of the various CD4⁺CD32 subsets, we quantified HIV-1 DNA in each population (total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high}, Fig. 1b) using droplet digital PCR (ddPCR), as described in the Methods. We found no evidence of HIV-1 DNA enrichment in the CD4⁺CD32^{high} fraction. We observed no significant difference in HIV-1 DNA between any populations and the CD4⁺CD32^{high} T cell population ($P = 0.28$). In fact, levels of HIV-1 DNA in the CD4⁺CD32^{high} T cell subsets isolated from nine participants was at the assay limit of detection (Fig. 1b). After correction for cell input in the CD4⁺CD32^{high} fraction, as estimated DNA values, we saw no evidence for HIV-1 DNA enrichment (open symbols in Extended Data Fig. 1b).

We then compared the relative frequency of the CD4⁺CD32^{high} T cell populations and the viral replicative capacity (IUPM values) per participant, but no relationship between the two parameters was observed (Fig. 1c). All values have been tabulated in Extended Data Table 2.

The HIV-1 reservoir largely resides in quiescent CD4⁺ T cells^{2,3}. Therefore, we sought to confirm the activation status of the CD4⁺ T cell populations by measuring the frequency of the activation markers CD69, CD25 and HLA-DR on CD4⁺ T cell subsets from all

participants. We found that the CD4⁺CD32^{high} T cells were highly activated compared to the CD4⁺CD32^{neg} T cells ($P < 0.0001$). Notably, among the activation markers, HLA-DR was particularly enriched,

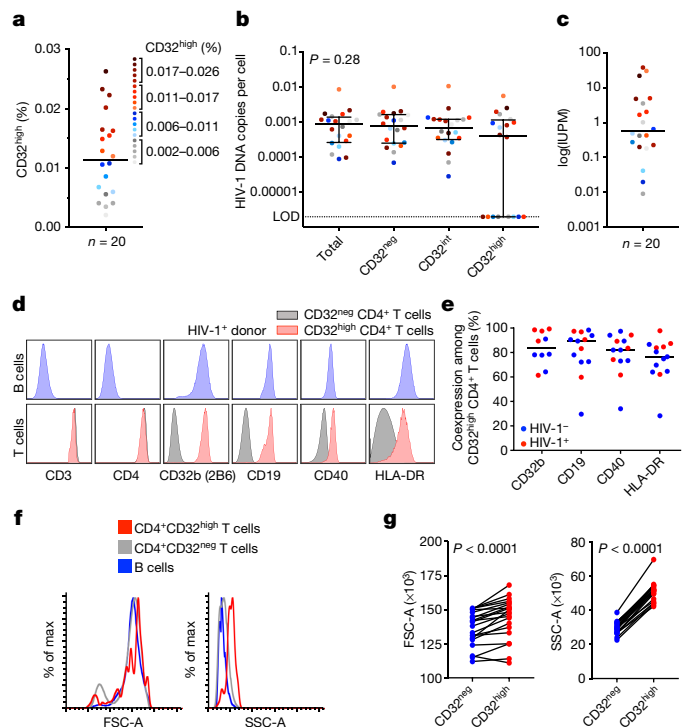


Fig. 1 | CD32-expressing CD4⁺ T cells are not enriched in HIV-1 DNA and express markers of B cell origin. **a–c**, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells from PBMCs of ART-suppressed, HIV-1-infected patients ($n = 20$) were sorted, and HIV-1 DNA was measured by ddPCR. **a**, Dividing the frequency (in percentage) of CD4⁺CD32^{high} T cells from all participants into quartiles, the values are shown as below or above the median. **b**, DNA copies per cell in sorted subsets of total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells are shown, with median and interquartile range (IQR). P value determined by Kruskal–Wallis test. LOD, limit of detection. **c**, IUPM in CD4⁺ T cells of each participant is shown in the colour corresponding to its frequency of CD4⁺CD32^{high} cells in panel **a**. **d**, **e**, CD32^{neg} and CD32^{high} (identified using FUN-2) CD4⁺ T cells from human PBMCs were assessed by flow cytometry for the expression of CD32b (2B6 antibody), CD19, CD40 and HLA-DR and compared to B cells (CD3⁺CD14⁺CD19⁺ lymphocytes). **d**, Representative flow cytometry results per cell antigen levels on B cells (top, blue histograms) and on CD32^{neg} and CD32^{high} CD4⁺ T cells (bottom, grey and red histograms, respectively) from PBMCs from an HIV-1⁺ participant. **e**, Frequency of CD4⁺CD32^{high} T cells staining positive for CD32b (2B6), CD19, CD40 or HLA-DR from HIV-1⁺ ($n = 5$) and HIV-1[−] ($n = 5–8$) human donor PBMC samples. Bars denote median values. **f**, Representative histograms of the FSC-A and SSC-A of B cells and CD32^{neg} and CD32^{high} CD4⁺ T cells from PBMCs of an HIV-1⁺, ART-suppressed participant sorted on a BD FACSAria II. **g**, Comparisons of the median FSC-A and SSC-A values between CD32^{neg} and CD32^{high} CD4⁺ T cell subsets from HIV-1⁺, ART-suppressed participants ($n = 20$). P values were determined using a paired t -test.

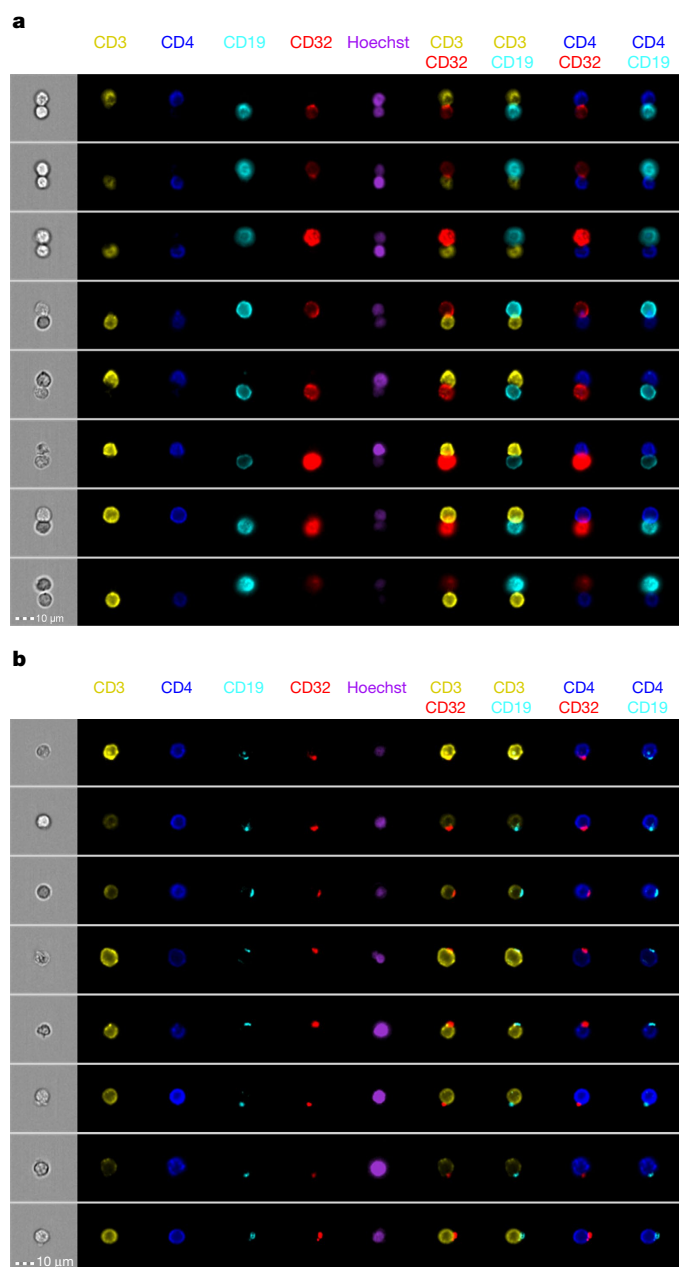


Fig. 2 | Flow cytometry imaging of sorted CD32-expressing T cells. **a**, Representative bright-field and pseudo-colour fluorescence images of T–B cell conjugates found in the CD32^{high} CD4⁺ T cell population sorted from PBMCs from HIV-1⁺, ART-suppressed participants, and imaged using Amnis technology. **b**, Representative images of punctate CD32 staining found on single T cells in the CD32^{high} and CD32^{int} population sorted from HIV-1⁺, ART-suppressed participant PBMCs.

marking approximately 75% of all CD4⁺CD32^{high} cells (median 74%), compared to CD4⁺CD32^{neg} cell populations (median 1.4%) (Extended Data Fig. 1c, $P < 0.0001$).

Two CD32 isoforms (CD32a and CD32b) are known to be expressed among all antigen presenting cells (APCs), but not typically on T cells. Therefore, we sought to exclude any APCs as potential contaminants of flow cytometry sorting. We evaluated the co-expression of lineage markers for all major CD32-bearing cells including monocytes, B cells, dendritic cells, granulocytes and natural killer cells. As expected, all CD32⁺ T cells expressed high amounts of CD3 and CD4 (Fig. 1d). However, we found that most CD4⁺CD32^{high} T cells from HIV-1⁺

patients, and also from healthy donors, co-expressed several B cell markers including CD19, CD40 and HLA-DR (Fig. 1d, e, Extended Data Fig. 2a). Notably, the B cell antigens found on CD4⁺CD32^{high} T cells were present at similar cell-surface densities as detected on bona fide B cells (Fig. 1d, e, Extended Data Fig. 2a).

The demonstration that the CD4⁺CD32^{high} fraction seen in HIV-1⁺ patients was marked with several B cell antigens and was similarly present in naive donors led us to investigate the origin of these B cell markers on a CD4⁺ T cell. Several reports have shown that B cells exclusively express the CD32b isoform⁴. The FUN-2 antibody clone used by Descours et al.¹ cannot distinguish between the CD32a and CD32b isoforms. Therefore, we used the monoclonal antibody clone 2B6 that has been reported to exclusively bind to CD32b^{4,5}. After co-staining PBMCs from HIV-1⁺ and HIV-1⁻ individuals with both the FUN-2 and the 2B6 antibodies, we found that all CD4⁺CD32^{high} T cells were marked only by the CD32b isoform and not by CD32a (Fig. 1d, e, Extended Data Fig. 2b), indicating that B cells are the origin of the CD32b antigen that marks the CD4⁺CD32^{high} T cells.

We sought to confirm this by determining whether *CD32A* (also known as *FCGR2A*) or *CD32B* (*FCGR2B*) mRNA was endogenously produced in the CD4⁺CD32^{high} subsets. After isolating total cellular RNA from various sorted T cell subsets, we used established reverse transcription PCR (RT–PCR) primers and probes that are specific to the CD32a and CD32b isoforms, as described in the Methods. We found that sorted CD4⁺CD32^{high} T cells from four HIV-1-infected participants did not contain detectable levels of the *CD32A* isoform. However, the *CD32B* mRNA isoform was readily detected in CD4⁺CD32^{high} T cells isolated from two out of four HIV-1⁺ patients (Extended Data Fig. 2c). By additional RT–PCR analysis, we detected both *CD3G* and *CD19* transcripts in the same CD4⁺CD32^{high} T population, indicating that the CD32b marking the CD4⁺CD32^{high} T cells may be from B cells expressing cognate CD32b (Extended Data Fig. 2d).

Because this may require cell-to-cell interaction, we performed a back-gating analysis of our flow cytometry data and confirmed that all CD4⁺CD32^{high} populations were identified within single-cell gates (Supplementary Fig. 1b). However, post-hoc analysis comparing the forward and side scatter light pulse area (FSC–A and SSC–A, respectively) values between CD4⁺CD32^{neg} and CD4⁺CD32^{high} T cells showed that the CD4⁺CD32^{high} populations had both a significantly higher FSC–A ($P < 0.0001$) and SSC–A ($P < 0.0001$), suggesting that the CD4⁺CD32^{high} population may consist largely of cell doublets (Fig. 1f, g).

We next used Amnis imaging flow cytometry to visualize the sorted CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} cell populations directly. As expected, the CD4⁺CD32^{neg} and the CD4⁺CD32^{int} cell populations each consisted of more than 99% single cells. However, the CD4⁺CD32^{high} fraction contained a high frequency of cell doublets (mean value 94%) (Extended Data Fig. 3). Of these ‘doublets’, approximately 70% seemed to be coincident doublets, and 30% were conjugates of T and B cells (Fig. 2a and Extended Data Fig. 3b).

We observed no examples in which CD32 staining on T cells was distributed throughout the cell membrane, supporting the idea that the CD32 found in the CD4⁺CD32^{high} population is not the result of endogenous expression from CD4⁺ T cells. Of the instances in which CD32 was detected on a T cell in the CD4⁺CD32^{high} population, the staining was punctate and often co-localized with punctate CD19 staining (Fig. 2b), suggesting that CD32 was acquired via contact between B and T cells. We noted that the frequency of T cells with punctate CD32 staining was substantially higher in the sorted CD32^{int} population. Thus, sorting for CD4⁺ T cells with a ‘high’ surface density of CD32 results in the selective enrichment of contaminating T–B cell doublets. As shown in Supplementary Fig. 1, these doublets cannot be discerned by routine cytometric FSC and SSC singlet gating strategies.

In summary, using samples from 20 HIV-1-infected, ART-suppressed participants, our data contradict the assertion that CD32a is a marker of

the replication-competent viral reservoir. Although we did detect similar frequencies of CD4⁺CD32^{high} populations to Descours et al.¹, we found no difference in the total HIV-1 DNA content between CD4⁺ T cell populations including or excluding the CD32^{high} fractions (Fig. 1b).

Notably, the CD4⁺CD32^{high} population was highly activated. Previous studies that have evaluated CD32 expression on T cells suggest that it may be detected after activation^{6,7} and led us to believe that this population may be atypical compared to a quiescent population harbouring the HIV-1 reservoir^{2,3}.

Our additional findings are incongruent with CD32a marking the replication-competent reservoir in CD4⁺ T cells; our phenotyping and RT-PCR experiments indicate that it is the CD32b isoform that marks the CD4⁺CD32^{high} cells (Fig. 1d, e, Extended Data Fig. 2b–d). This finding, combined with the demonstration that this cell population is found in uninfected individuals, conflicts with the assertion of Descours et al.¹ that CD32a is upregulated after the establishment of viral latency. Recent reports have corroborated the absence of CD32a transcripts in reactivated, clonal HIV-1-infected CD4⁺ T cells⁸.

The surface density of CD32b (and other B cell markers) on the CD4⁺CD32^{high} population was observed at similar densities to that on B cells. These data, combined with the post-hoc analysis, suggests that this population may be largely comprised of doublets. Direct interrogation of the CD4⁺CD32^{high} population via Amnis imaging confirmed that this population consisted largely of contaminating doublets; either co-incident events or cell-to-cell conjugates (Fig. 2a).

We demonstrate that the mechanism by which the CD32b isoform labels the CD4⁺CD32^{high} populations is through the direct interaction of CD4⁺ T and B cells, and possible trogocytotic transfer of B cell antigens to T cells, as observed in the CD4⁺CD32^{int} population (Fig. 2b). This may explain the transfer or membrane painting of antigens such as CD32b, CD40 and HLA-DR, among other markers^{9–11}. Not only have cell-to-cell membrane transfers been shown to occur commonly *in vivo* during viral infections, but such transfers largely occur on activated cells¹². Membrane-bound Fcγ receptors, including CD32b, are known to be extracted from APCs and then transferred to T cells, and serve as a surrogate of recent T cell and APC interactions¹³. Our demonstration of T–B cell conjugates in the CD4⁺CD32^{high} population and high levels of single cells in the CD4⁺CD32^{int} population support this notion (Fig. 2a, b).

Collectively, our findings confirm that selectively sorting for T cells with a high surface density of CD32 results in the enrichment of T–B cell doublet contaminants, which cannot be discerned by routine gating strategies. The true isoform, CD32b, that marks the CD4⁺CD32^{high} population is probably indicative of dynamic CD4⁺ T cell interaction with B cells, rather than a marker of the HIV-1 reservoir^{14,15}.

We thank S. Mordecai for Amnis technical expertise, and acknowledge support from NIAID grants AI091514, AI122942, AI127089 and AI131365 awarded to J.B.W. Support was also provided by the NIAID awarded Martin Delaney Collaboratory ‘BELIEVE’ grant AI126617, co-funded by NIDA, NIMH and NINDS awarded to D.F.N.

Methods

HIV-1⁺ participants were recruited through: The Maple Leaf Medical clinic in Toronto, Canada; The HIV Eradication and Latency (HEAL) cohort of Brigham and Women's and Massachusetts General Hospital; The Whitmann Walker Clinic in Washington, DC; or the Hospital of the University of Pennsylvania. The study was approved by the University of Toronto, The University of Pennsylvania and George Washington University ethics committees and according to the protocol approved by the Partners Human Research Committee and Institutional Review Board (IRB). Written informed consent was obtained from each participant.

The percentage of CD32⁺ (clone FUN-2) CD4⁺ T cells was measured in samples from study participants. Both CD32⁺ and CD32^{neg} CD4⁺ T cells were sorted and viral DNA was measured using ddPCR. The analysis of cell lineage markers by flow cytometry and RT-PCR was also conducted. Flow cytometry sorts from PBMCs used in HIV-1 DNA analyses were performed on cell subsets and assessed using Amnis imaging flow cytometry.

Data availability. All data and reagents are available from the corresponding author upon request.

Christa E. Osuna¹, So-Yon Lim¹, Jessica L. Kublin¹, Richard Apps², Elsa Chen¹, Talia M. Mota³, Szu-Han Huang³, Yanqin Ren³, Nathaniel D. Bachtel³, Athe M. Tsibris⁴, Margaret E. Ackerman⁵, R. Brad Jones³, Douglas F. Nixon³ & James B. Whitney^{1,6*}

¹Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. ²Center for Human Immunology, National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA. ³Division of Infectious Diseases, Weill Department of Medicine, Weill Cornell Medical College, New York, NY, USA. ⁴Brigham and Women's Hospital, Boston, Massachusetts Harvard Medical School, Boston, MA, USA. ⁵Thayer School of Engineering, Dartmouth College, Hanover, NH, USA. ⁶Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA. *e-mail: jwhitne2@bidmc.harvard.edu

Received: 11 August 2017; Accepted: 24 May 2018;
Published online 19 September 2018.

- Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
- Chun, T. W. et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387**, 183–188 (1997).
- Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
- Veri, M. C. et al. Monoclonal antibodies capable of discriminating the human inhibitory Fcγ-receptor IIB (CD32B) from the activating Fcγ-receptor IIA (CD32A): biochemical, biological and functional characterization. *Immunology* **121**, 392–404 (2007).
- Boruchov, A. M. et al. Activating and inhibitory IgG Fc receptors on human DCs mediate opposing functions. *J. Clin. Invest.* **115**, 2914–2923 (2005).
- Engelhardt, W., Matzke, J. & Schmidt, R. E. Activation-dependent expression of low affinity IgG receptors FcγRII(CD32) and FcγRIII(CD16) in subpopulations of human T lymphocytes. *Immunobiology* **192**, 297–320 (1995).
- Sandilands, G. P. et al. Differential expression of CD32 isoforms following alloactivation of human T cells. *Immunology* **91**, 204–211 (1997).
- Cohn, L. B. et al. Clonal CD4⁺ T cells in the HIV-1 latent reservoir display a distinct gene profile upon reactivation. *Nat. Med.* **24**, 604–609 (2018).
- Cone, R. E., Sprent, J. & Marchalonis, J. J. Antigen-binding specificity of isolated cell-surface immunoglobulin from thymus cells activated to histocompatibility antigens. *Proc. Natl Acad. Sci. USA* **69**, 2556–2560 (1972).
- Hwang, I. et al. T cells can use either T cell receptor or CD28 receptors to absorb and internalize cell surface molecules derived from antigen-presenting cells. *J. Exp. Med.* **191**, 1137–1148 (2000).
- Wetzel, S. A., McKeithan, T. W. & Parker, D. C. Peptide-specific intercellular transfer of MHC class II to CD4⁺ T cells directly from the immunological synapse upon cellular dissociation. *J. Immunol.* **174**, 80–89 (2005).
- Rosenits, K., Keppler, S. J., Vucikujia, S. & Aichele, P. T cells acquire cell surface determinants of APC via *in vivo* trogocytosis during viral infections. *Eur. J. Immunol.* **40**, 3450–3457 (2010).
- Daubeuf, S. et al. Preferential transfer of certain plasma membrane proteins onto T and B cells by trogocytosis. *PLoS One* **5**, e8716 (2010).
- Garside, P. et al. Visualization of specific B and T lymphocyte interactions in the lymph node. *Science* **281**, 96–99 (1998).
- Okada, T. et al. Antigen-engaged B cells undergo chemotaxis toward the T zone and form motile conjugates with helper T cells. *PLoS Biol.* **3**, e150 (2005).

Author contributions D.F.N. and J.B.W. designed the studies. R.B.J., R.A., E.C., Y.R., N.D.B., C.E.O., R.T. and S.Y.L. led the virology assays. S.H.H., D.C., J.L.K., M.A. and C.E.O. led the immunology assays. J.B.W. led the studies and wrote the paper with all co-authors.

Competing interests Declared none.

Additional information

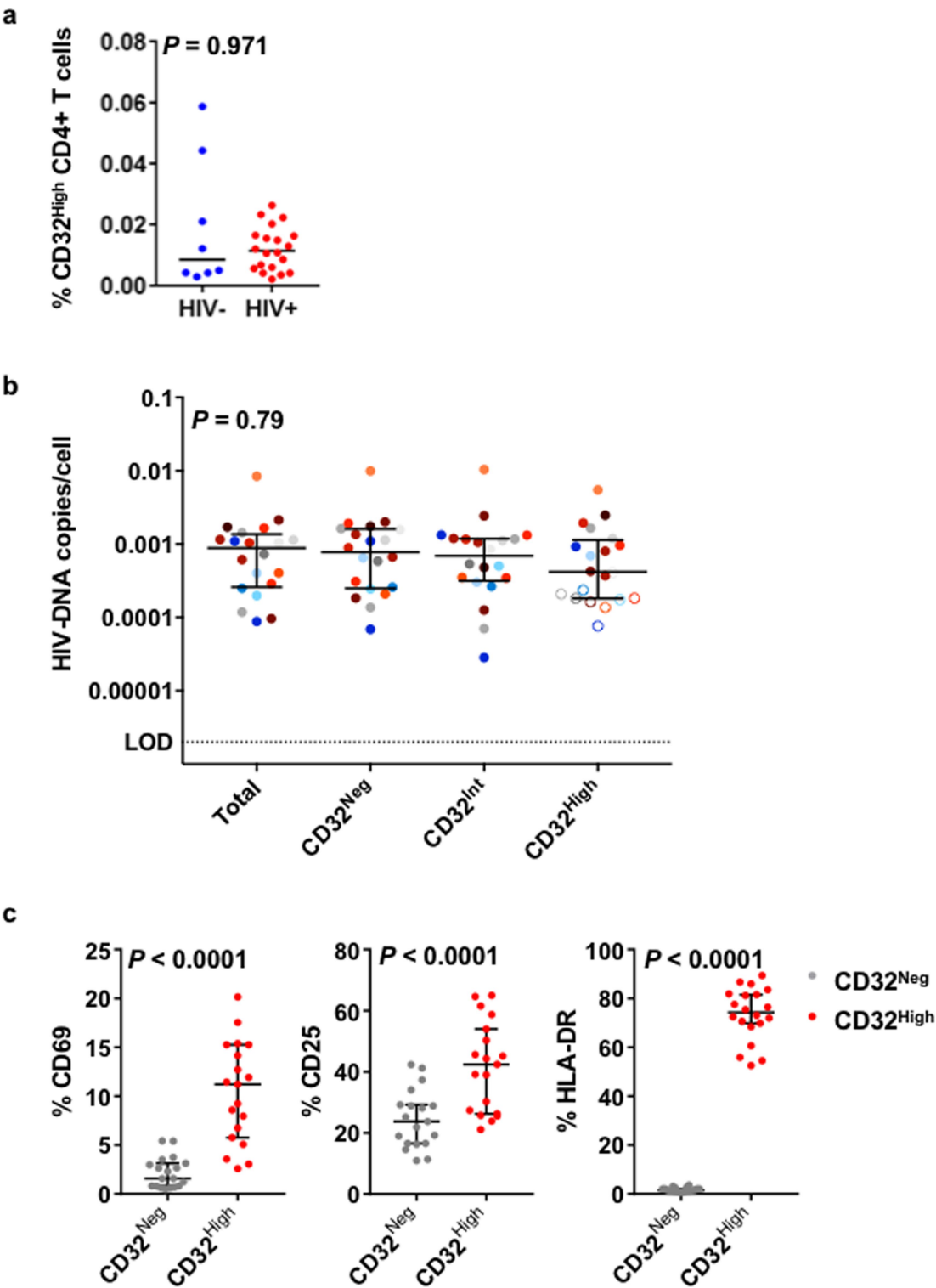
Extended data accompanies this Comment.

Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

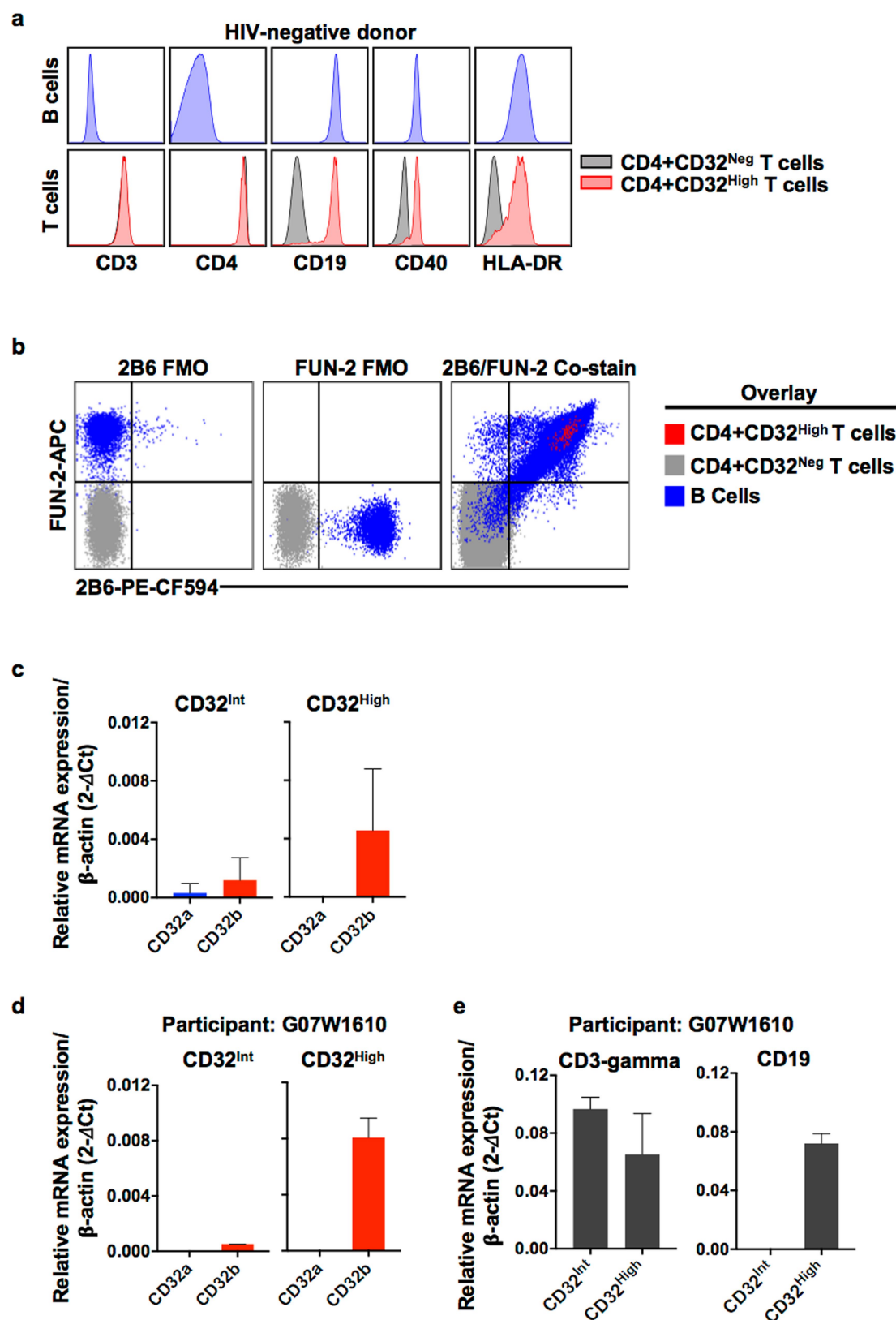
Correspondence and requests for materials should be addressed to J.B.W.

<https://doi.org/10.1038/s41586-018-0495-2>



Extended Data Fig. 1 | Frequency and activation status of CD32-expressing CD4⁺ T cells and their HIV-1 DNA content. **a**, The frequency of CD32^{high} CD4⁺ T cells was measured by flow cytometry in PBMCs from ART-suppressed, HIV-1⁺ ($n = 20$) and HIV-1⁻ ($n = 8$) donors. Bars denote median values. P values were determined by a Mann–Whitney test. **b**, DNA copies per cell in sorted subsets of total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells are shown with median values and the IQR. The results are shown as either the actual HIV-1 DNA

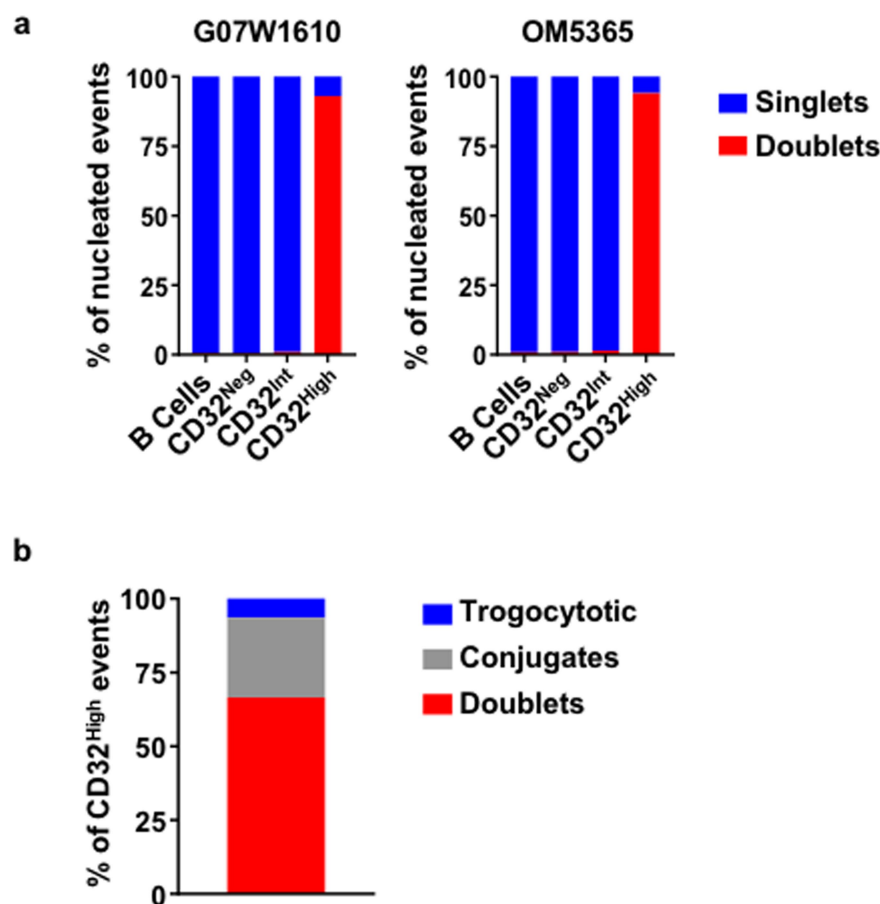
copies per million cells (filled symbols) or as estimated values calculated using the LOD and applied to the number of cells when the DNA input did not reach the threshold (open symbols). P values were determined by a Kruskal–Wallis test. **c**, The percentage of CD69, CD25 and HLA-DR expression was measured by flow cytometry on CD32^{neg} and CD32^{high} (FUN-2) CD4⁺ T cells from PBMCs from HIV-1⁺ participants ($n = 20$). Error bars show the median and IQR. P values were determined by Wilcoxon matched-pairs signed rank tests.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Detection of B cell proteins and mRNA in CD32-expressing CD4⁺ T cells. **a**, CD32^{neg} and CD32^{high} (FUN-2) CD4⁺ T cells from human PBMCs were assessed by flow cytometry for the expression of CD19, CD40 and HLA-DR, and compared to B cells (CD3⁻ CD14⁻ CD19⁺ lymphocytes). Representative flow cytometry results of per cell antigen levels on B cells (top, blue histograms) and CD32^{neg} and CD32^{high} CD4⁺ T cells (bottom, grey and red histograms, respectively) from an HIV-1⁻ donor. **b**, Representative CD32b staining of PBMCs from an HIV-1⁺, ART-suppressed participant. PBMCs were stained with an optimized concentration of the 2B6 monoclonal anti-CD32b antibody, followed by an antibody cocktail that included the FUN-2 monoclonal pan-CD32 antibody, as described in the Methods. Shown are the 2B6 and FUN-2

fluorescence minus one (FMO) antibody cocktail-stained samples and a sample co-stained with 2B6 and FUN-2. **c**, **d**, CD32 mRNA expression levels in CD4⁺CD32⁺ subsets. **c**, The relative expression of CD32A and CD32B mRNA isoforms in sorted CD4⁺CD32^{int} and CD4⁺CD32^{high} subsets from HIV-1⁺, ART-suppressed participants ($n = 4$). **d**, mRNA expression of CD32A and CD32B from patient G07W1610. **e**, T and B cell lineage-specific mRNA transcripts in sorted CD4⁺CD32⁺ subsets from participant G07W1610. Relative mRNA expression of target genes was normalized to *ATCB* using the comparative C_t method. Results are mean \pm s.d. of each value from each participant ($n = 4$; **c**), or from values generated from two separate experiments using samples from the same patient (**d**).



Extended Data Fig. 3 | Doublet composition of the sorted CD4⁺CD32^{high} T cells. Sorted B cells and CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells from an HIV-1⁺, ART-suppressed participant were analysed using an Amnis imaging cytometer. Singlets and doublets were quantified using the aspect ratio and nuclear staining. **a**, The proportion of total singlet and doublet events among total nucleated

cells detected on the Amnis cytometer in each sorted population was determined, and is shown as individual composite bar graphs for two patients (G07W1610 and OM5365). **b**, A composite bar graph of the proportion of conjugates, doublets and trogocytotic events that comprised the sorted CD4⁺CD32^{high} population ($n = 2$).

BRIEF COMMUNICATIONS ARISING

Extended Data Table 1 | Viral suppression of 20 HIV-1-infected participants on ART

Cohort	Participant ID	Date of Initial Suppression (MM/YY)	Length of suppression (yrs)
HEAL	HEAL-009	3/14	3
	HEAL-019	8/09	8
	HEAL-020	3/08	9.3
	HEAL-034	11/05	11.5
	HEAL-053	11/16	1
	HEAL-055	11/00	17
Maple Leaf	CIRC0024	6/98	17.0
	CIRC0133	7/08	7.0
	CIRC0196	4/14	1.2
	OM5011	11/08	6.6
	OM5148	1/08	7.5
	OM5162	9/04	10.8
	OM5203	3/12	3.3
	OM5334	7/14	0.9
	OM5365	3/08	7.3
WWH	WWH-B001	7/11	6.4
	WWH-B005	12/17	0.3
	WWH-B008	11/14	3.1
	WWH-B011	11/11	6
UPenn	G07W1610	10/05	11.8

BRIEF COMMUNICATIONS ARISING

Extended Data Table 2 | CD4⁺CD32^{high} subset proportions and HIV-1 DNA compared to total CD4⁺ and CD32^{neg} CD4⁺ T cells

Participant ID	CD32 ^{High}		HIV-DNA enrichment			
	% in total CD4	Absolute cell count	HIV-DNA copies/cell ¹	CD32 ^{High} /CD4 total ²	CD32 ^{High} /CD32 ^{Neg2}	CD32 ^{Neg} /CD4 total
HEAL-009	0.007	11,427	>0.000002	0.010	0.008	1.236
HEAL-019	0.002	4,911	>0.000002	0.002	0.001	1.500
HEAL-020	0.011	8,482	>0.000002	0.008	0.008	1.036
HEAL-034	0.022	8,238	0.000426	0.199	0.212	0.937
HEAL-053	0.004	9,544	>0.000002	0.017	0.015	1.161
HEAL-055	0.011	21,806	0.00037	0.604	0.555	1.088
CIRC0024	0.015	26,200	>0.000002	0.023	0.029	0.782
CIRC0133	0.017	14,602	>0.000002	0.005	0.010	0.517
CIRC0196	0.006	5,935	0.000694	1.722	1.066	1.615
OM5011	0.008	8,862	0.001942	1.871	2.187	0.855
OM5148	0.004	8,788	0.001191	1.043	1.049	0.994
OM5162	0.016	7,133	0.000923	0.842	0.842	1.000
OM5203	0.026	12,254	>0.000002	0.021	0.011	1.911
OM5334	0.016	6,275	0.000959	0.579	0.499	1.160
OM5365	0.006	11,027	>0.000002	0.003	0.003	0.803
WWH-B001	0.020	10,922	>0.000002	0.007	0.006	1.076
WWH-B005	0.013	5,964	0.00547	0.650	0.550	1.182
WWH-B008	0.023	6,464	0.000805	0.696	0.595	1.169
WWH-B011	0.004	5,953	0.001653	1.154	1.016	1.135
G07W1610	0.012	7,984	0.002482	1.452	1.411	1.029
Median	0.012	8,635	0.000398	0.389	0.356	1.082

¹Values below the LOD (2 copies per 10⁶ cells) are shaded in grey.

²To calculate HIV-1 enrichment, 0.000002 was used for all values below the LOD.

Descours et al. reply

REPLYING TO L. Pérez et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0493-4> (2018); C. E. Osuna et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0495-2> (2018); L. N. Bertagnolli et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0494-3> (2018)

In our previous work¹, we used an in vitro model of HIV-infected unstimulated CD4 T cells to identify CD32 as a candidate marker of HIV⁺ resting CD4 T cells in vitro, and a subset of HIV⁺ total CD4 T cells containing replication-competent viruses in individuals that underwent anti-retroviral therapy (ART). Of note, we did not explore the transcriptional status of hosted viruses (latent or active) ex vivo, nor the activation state of these cells (quiescent or activated)¹. In the accompanying Comments^{2–4}, colleagues attempted to reproduce these findings. They present experiments that support the following conclusions: (1) the isolation of the CD32⁺ CD4 T cell population results from artefacts caused by the flow cytometry sorting method^{2,3}, and (2) the sorted CD32 CD4 T cell population is not enriched in HIV nor in replication-competent proviral DNA^{2–4}. Here, we formulate two questions that mirror the major issues raised by these three Comments^{2–4} and discuss their results in the context of our previous report¹ and more recently published studies.

Is there any evidence that a CD4 T cell can express CD32 in the context of HIV infection? This question is raised by both Osuna et al.² and Pérez et al.³. A recent report⁷, using in situ hybridization (which avoids the criticism of artefacts caused by flow cytometry sorting), showed that HIV-1 RNA co-localized with CD32A (also known as FCGR2A) RNA in 90% of examined cells in B cell follicles from four individuals. Because HIV primarily targets CD4 T cells, these data may support the ability of a CD4 T cell to upregulate CD32 mRNA transcription after infection in vivo. Three independent groups have identified CD32 as being expressed by latently or productively infected CD4 T cells in vitro^{1,5–7}. These models generated and analysed a substantial percentage of HIV-infected CD4 cells. Thus, any marker that is usually not expressed by CD4 T cells but that is detected at the surface of these cells after infection is unlikely to result from biased analyses of cellular doublets, as could be the case when working on rare events from ex vivo samples^{2,3}. Instead, these data suggest that transcriptional regulation leading to the expression of CD32 mRNA and protein can probably occur after in vitro and in vivo infection of a single CD4 T cell.

Does the CD32 CD4 T cell subset contribute to viral persistence under treatment? All three of the accompanying Comments^{2–4} indicate that CD32 CD4 T cells are not enriched for HIV DNA in blood. Recent work suggests, however, that in some virally suppressed HIV-infected individuals, CD32 CD4 T cells were enriched in HIV DNA, although to a lesser extent than we reported⁸. Notably, this question has been recently addressed in tissues, and results seem to be less contrasted than in blood^{7,9,10}. More importantly, they revealed functional properties of these reservoir cells that have not been previously explored^{7,9,10}. As discussed above, a recent report⁷ found that within the B cell follicles of virally suppressed HIV-infected individuals, most of the cells containing HIV RNA and persisting despite treatment were found to express CD32A RNA⁷. This result seems to be in line with other data¹⁰ that indicate that T follicular helper cells, primarily found in these territories, were enriched for HIV DNA and RNA when expressing CD32¹⁰, although at a lower extent than our previous findings¹. In non-lymphoid rectal tissue, CD4 T cells expressing CD32 were also enriched

for both HIV DNA and RNA⁹. Notably, the co-expression of CD32 and HIV RNA reported in these two publications^{9,10} suggests that CD32 marks transcriptionally active infected cells rather than latent cells. Together, these reports support the ability of CD32 to identify a subset of persistent HIV-infected CD4 T cells and suggest that they could contribute to viral persistence under ART in vivo.

In conclusion, we believe that rather than completely ruling out the relevance of CD32 for the identification of a subset of infected cells in vivo and their contribution to HIV persistence, the whole literature, including the three accompanying Comments^{2–4}, opens new technical challenges and questions that we should solve in the near future.

Benjamin Descours, Gael Petitjean and Monsef Benkirane are solely responsible for this Reply. The contributions of the remaining authors from the original Letter¹ were limited to recruiting patients or performing analysis on blinded samples, and thus only Descours, Petitjean and Benkirane have authored this Reply.

Benjamin Descours¹, Gael Petitjean¹ & Monsef Benkirane^{1*}

¹Institut de Génétique Humaine, Laboratoire de Virologie Moléculaire, UMR9002, CNRS, Université de Montpellier, Montpellier, France.

*e-mail: monsef.benkirane@igh.cnrs.fr

1. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
2. Osuna, C. E. et al. Evidence that CD32a does not mark the HIV-1 latent reservoir. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0495-2> (2018).
3. Pérez, L. et al. Conflicting evidence for HIV enrichment in CD32⁺ CD4 T cells. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0493-4> (2018).
4. Bertagnolli, L. N. The role of CD32 during HIV-1 infection. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0494-3> (2018).
5. Iglesias-Ussel, M., Vandergeeten, C., Marchionni, L., Chomont, N. & Romero, F. High levels of CD2 expression identify HIV-1 latently infected resting memory CD4⁺ T cells in virally suppressed subjects. *J. Virol.* **87**, 9148–9158 (2013).
6. Grau-Expósito, J. et al. A Novel single-cell FISH-flow assay identifies effector memory CD4⁺ T cells as a major niche for HIV-1 transcription in HIV-infected patients. *MBio* **8**, e00876-17 (2017).
7. Abdel-Mohsen, M. et al. CD32 is expressed on cells with transcriptionally active HIV but does not enrich for HIV DNA in resting T cells. *Sci. Transl. Med.* **10**, eaar6759 (2018).
8. Martin, G. E. et al. CD32-expressing CD4 T cells are phenotypically diverse and can contain proviral HIV DNA. *Front. Immunol.* **9**, 928 (2018).
9. Hogan, L. E. et al. Increased HIV- transcriptional activity and infectious burden in peripheral blood and gut-associated CD4⁺ T cells expressing CD30. *PLoS Pathog.* **4**, e006856 (2018).
10. Noto, A., Procopio, F., Corpataux, J. M. & Pantaleo, G. CD32⁺PD1⁺ Tfh cells are the major HIV reservoir in long-term art-treated individuals. *J. Virol.* <https://doi.org/10.1128/JVI.00901-18> (2018).

Author contributions B.D., G.P. and M.B. wrote the manuscript.

Competing interests Declared none.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.B.

<https://doi.org/10.1038/s41586-018-0496-1>

Conflicting evidence for HIV enrichment in CD32⁺ CD4 T cells

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

Descours and colleagues¹ reported a marked enrichment for HIV among CD32a⁺ CD4 T cells in people receiving anti-retroviral therapy (ART). This tiny CD32a⁺ population (0.012% of all blood CD4 T cells) contained a median of 0.56 HIV DNA genomes per cell, and accounted for 26.8–86.3% of HIV DNA in CD4 T cells, thus suggesting that targeting CD32a⁺ CD4 T cells might help to clear HIV reservoirs in vivo. Here, we report our unsuccessful attempts to confirm these findings. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0496-1> (2018).

We first used fluorescence-activated cell sorting (FACS) to sort CD4 T cells with high, intermediate and low levels of CD32 staining (CD32^{hi}, CD32^{int} and CD32^{lo}, respectively) from 10 individuals with chronic HIV infection who were receiving ART (mean duration, 8.8 years; range, 2.7–15). We used cell-staining reagents and gating techniques that matched those used by Descours et al.¹ (see Supplementary Methods and Extended Data Fig. 1). As shown in Fig. 1a, we detected no enrichment for HIV DNA in the CD32^{hi} or CD32^{int} CD4 T cells. Moreover, the CD32^{hi} and CD32^{int} subsets combined accounted for no more than 3% of all HIV DNA copies within circulating CD4 T cells in any of the 10 study participants (Fig. 1b). Post-sort flow cytometry of CD32^{hi} and CD32^{int} populations showed heterogeneous patterns that suggested the formation of T cell–B cell or T cell–monocyte conjugates as the origin of most CD32^{hi} or CD32^{int} CD4 T cells, with separation of these conjugates during sorting (Extended Data Fig. 2).

To rule out the possibility that we had inadvertently obtained false negative results either by excluding HIV-infected, CD32⁺ CD4 T cells using tight light scatter gates or by failing to exclude non-T-cell contaminants, we performed parallel sorts on the same 10 samples using an alternative gating scheme. We used a more inclusive light scatter gate as well as markers for B cells, monocytes, dendritic cells and natural killer cells (Extended Data Fig. 3). Events that were CD3⁺ were separated into fractions that were positive for B cell markers (T–B), positive for one or more other non-CD4-T-cell markers (T–other), or negative for all of these, positive for CD4, and CD32^{hi}, CD32^{int} or CD32^{lo}. Neither CD32^{hi} nor CD32^{int} CD4 T cells were enriched for HIV DNA (Fig. 2a). Similarly, we detected no enrichment for HIV DNA in the T–B and

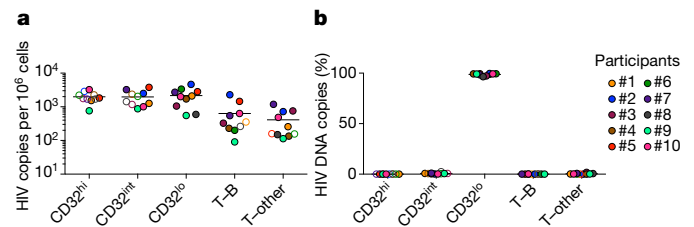


Fig. 2 | Levels of HIV DNA in CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cells, sorted using alternative gating. The samples from Fig. 1 were sorted using alternative gating in which T cells bearing markers of B cells (T–B) or other non-CD4-T-cell lineages (T–other) were first collected in separate tubes. **a**, Copies of HIV DNA per million sorted cells. **b**, Percentages of all HIV DNA copies detected in blood cells that were detected within each subset, calculated by adjusting values in **a** for the relative proportions of these subsets in FACS data.

T–other populations (Fig. 2a). In each of the 10 participants, at least 96% of all HIV DNA copies occurred in conventional CD32^{lo} cells (Fig. 2b). Post-sort flow cytometry suggested that most events bearing both T-cell and non-CD4-T-cell markers again represented cell–cell conjugates, and also showed that most remaining CD32^{hi} CD4 T cells did not reproducibly show a high CD32 signal after sorting (Extended Data Fig. 4). This was in contrast to conventional CD32^{lo} cells, which were uniformly pure in post-sort analyses across participants. In a second group of four individuals whose peripheral blood mononuclear cells (PBMCs) were sorted without previous cryopreservation (Extended Data Fig. 5a), we again found no enrichment for HIV DNA based on CD32 expression (Extended Data Fig. 5b), and also observed that HIV DNA sequences in CD32⁺ CD4 T cells were genetically intermingled with HIV DNA sequences in other CD4 T cells (Extended Data Fig. 5c).

Overall, our studies showed no enrichment for HIV DNA in CD32⁺ CD4 T cells, and also raised questions about the source of the CD32 labelling on these cells. We propose that the CD32 expression associated previously with CD4 T cells could have arisen from adherent non-T-cells or cellular material bearing this marker, and that conjugates containing HIV-infected CD4 T cells could be differentially produced and/or recovered in different laboratories with different sample processing and FACS practices. It is important to acknowledge that these considerations do not explain the discrepancy between the Descours et al. study¹ and ours in the quantities of HIV DNA detected within CD3⁺CD4⁺CD32⁺ sorted material. Nevertheless, we wish to emphasize that our findings do not support targeting CD32 molecules on CD4 T cells in emerging HIV cure strategies.

Methods

Participant recruitment and informed consent were performed under Institutional Review Board (IRB)-approved protocols at the US National Institutes of Health (NIH). For FACS, whole PBMCs were stained with monoclonal antibodies matching those used by Descours et al.¹ (see Supplementary Methods) and sorted on a BD FACSARIA. To evaluate purity, a portion of each population was re-analysed on the flow cytometer after sorting. Virus DNA copies in sorted cells were enumerated by fluorescence-assisted clonal amplification². DNA recovery was quantified by albumin (*ALB*) quantitative PCR. Because the FUN-2 monoclonal antibody used

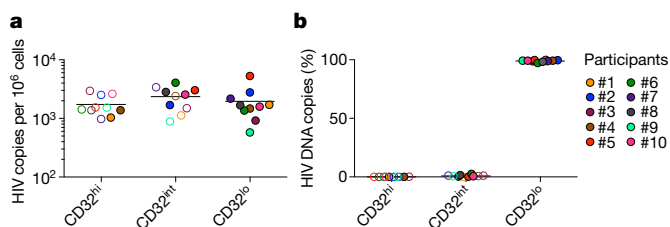


Fig. 1 | Levels of HIV DNA in CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cells, sorted from PBMCs of 10 ART-treated participants, as in Descours et al.¹ **a**, Copies of HIV DNA per million sorted cells. **b**, Percentages of all HIV DNA copies detected in blood CD4 T cells that were detected within each subset, calculated by adjusting values in **a** for the relative proportions of these subsets in FACS data. In all figures, horizontal bars denote median values, and open symbols indicate detection limits for measurements in which HIV DNA was not detected.

BRIEF COMMUNICATIONS ARISING

by Descours et al.¹ and in our study may recognize both CD32a and CD32b, we refer to cells staining with this monoclonal antibody as CD32⁺.

Data availability. All DNA sequences in this manuscript (analysed in Extended Data Fig. 5) have been deposited in GenBank under accession numbers MH080310–MH080572.

Liliana Pérez¹, Jodi Anderson², Jeffrey Chipman³, Ann Thorkelson², Tae-Wook Chun⁴, Susan Moir⁴, Ashley T. Haase⁵, Daniel C. Douek⁶, Timothy W. Schacker^{2,7} & Eli A. Boritz^{1,7*}

¹Virus Persistence and Dynamics Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. ²Division of Infectious Diseases, University of Minnesota, Minneapolis, MN, USA. ³Department of Surgery, University of Minnesota, Minneapolis, MN, USA. ⁴Laboratory of Immunoregulation, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD, USA. ⁵Department of Microbiology and Immunology, University of Minnesota, Minneapolis, MN, USA. ⁶Human Immunology Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. ⁷These authors jointly supervised this work: Timothy W. Schacker, Eli A. Boritz. *e-mail: boritze@mail.nih.gov

Received: 11 October 2017; Accepted: 20 March 2018;

Published online 19 September 2018.

1. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
2. Boritz, E. A. et al. Multiple origins of virus persistence during natural control of HIV infection. *Cell* **166**, 1004–1015 (2016).

Author contributions Data generation and analysis: L.P., J.A., T.W.S. and E.A.B. Study design and oversight: L.P., A.T.H., D.C.D., T.W.S. and E.A.B. Participant cohort and sample management: J.A., J.C., A.T., T.W.C., S.M. and T.W.S. Manuscript preparation: L.P., A.T.H., D.C.D., T.W.S. and E.A.B.

Competing interests Declared none.

Additional information

Extended data accompanies this Comment.

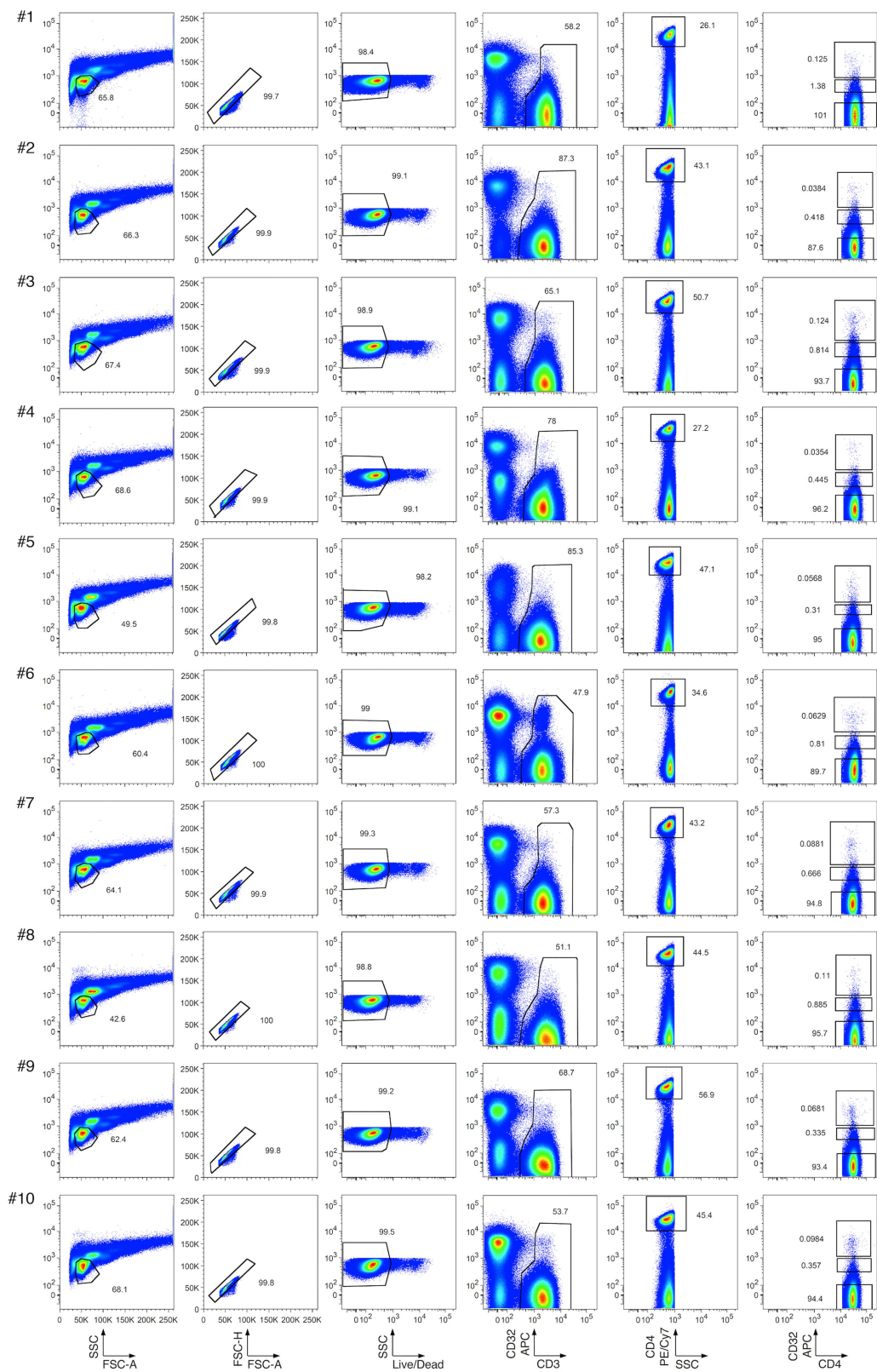
Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to E.A.B.

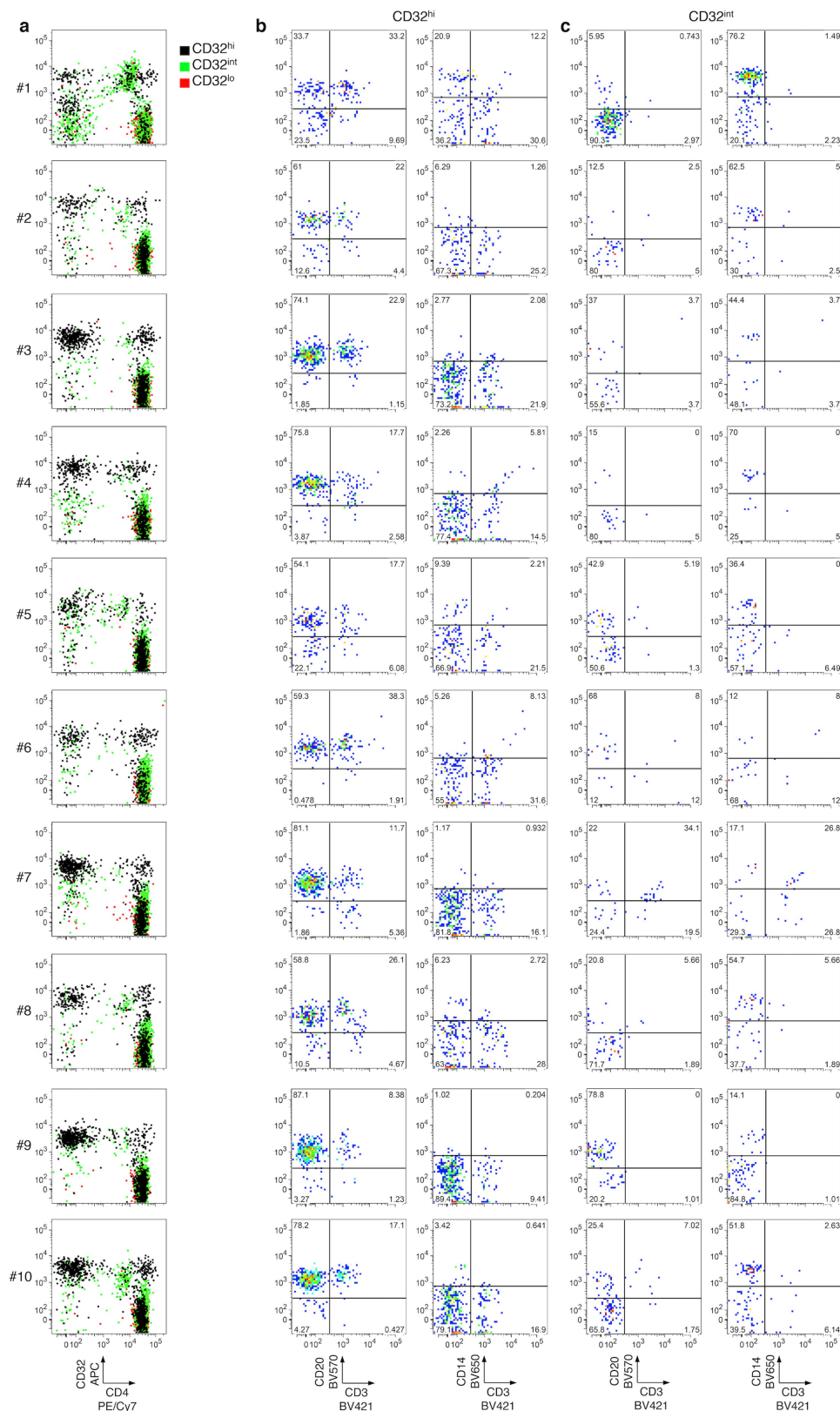
<https://doi.org/10.1038/s41586-018-0493-4>

BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 1 | Flow cytometry of CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations from PBMCs. Single lymphocytes (first two columns) that were viable (third column), CD3⁺ (fourth column), CD4⁺

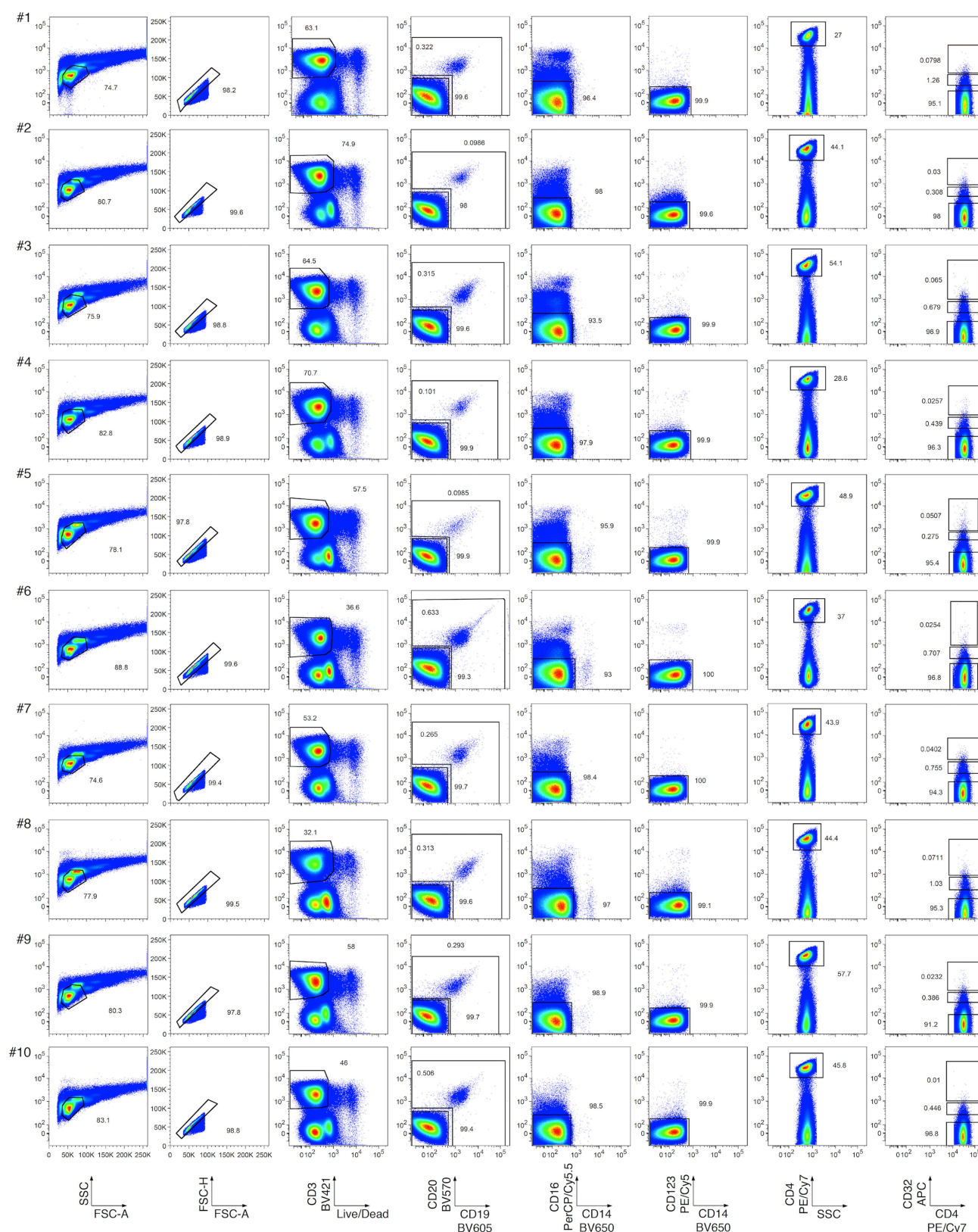
(fifth column), and CD32^{hi}, CD32^{int} or CD32^{lo} (sixth column) were sorted as described in Descours et al.¹.



Extended Data Fig. 2 | Post-sort flow cytometry of CD32⁺CD4⁺ subsets that were CD32^{hi}, CD32^{int} or CD32^{lo}. Cells were sorted as in Extended Data Fig. 1. **a**, Overlay plots of CD32 and CD4 expression by cells in CD32^{hi}, CD32^{int} and CD32^{lo} sorted populations. Note the heterogeneous pattern of cells from the CD32^{hi} and CD32^{int} populations. **b**, **c**, CD20,

CD14 and CD3 staining in the CD32⁺ cells from the CD32^{hi} (**b**) and the CD32^{int} (**c**) subsets. Note the large proportions of all CD32⁺ cells bearing surface markers consistent with B cells (CD20⁺CD3⁻) or monocytes (CD14⁺CD3⁻) after sorting.

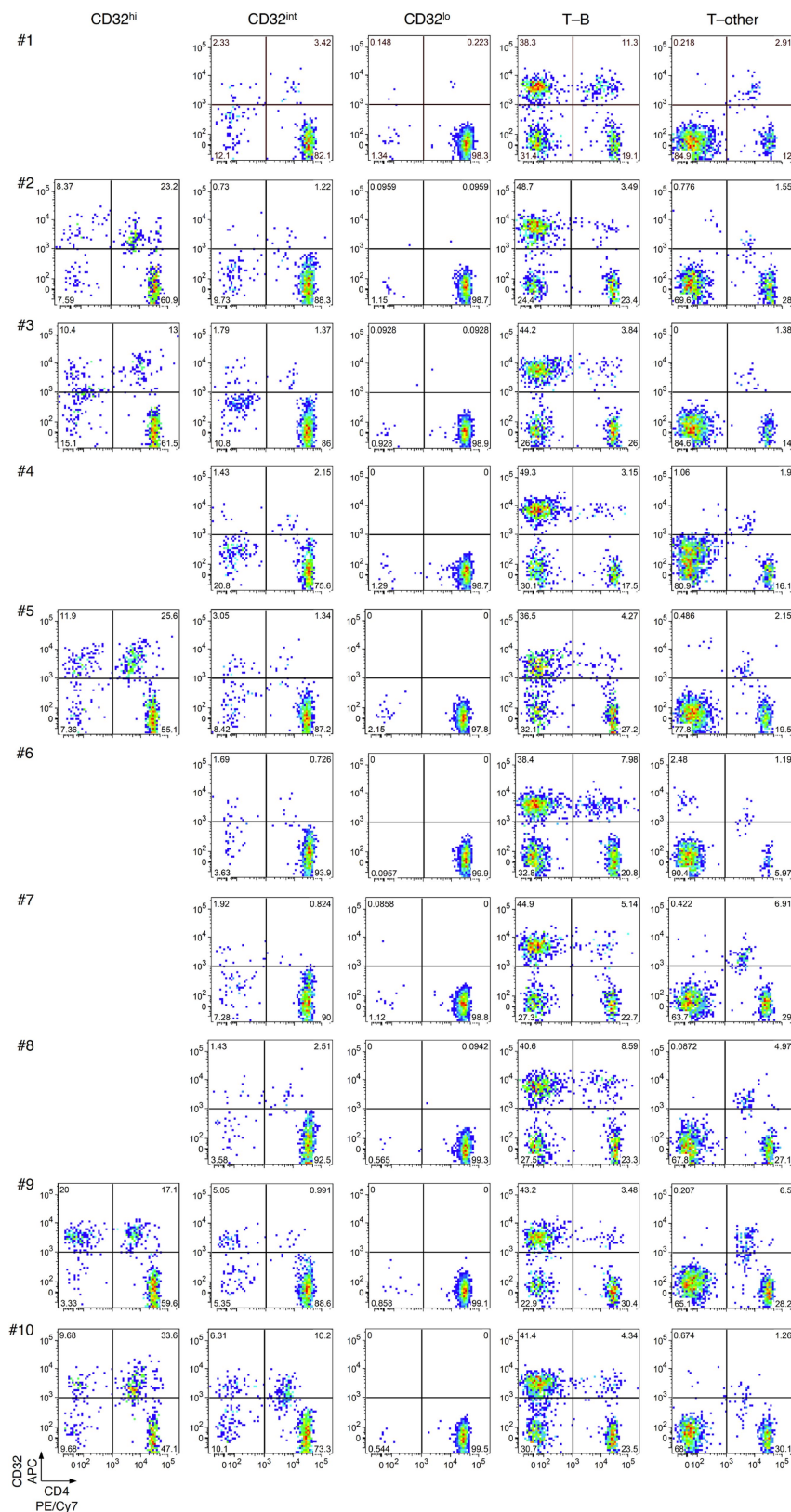
BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 3 | Flow cytometry of PBMCs sorted by alternative gating for CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations, as well as T cell populations bearing markers of B cells (T-B) or other non-CD4-T-cells (T-other). Cells in an inclusive light scatter gate consistent with either small lymphocytes or larger cells (first column) were enriched for single cells (second column). Within these gates, viable CD3⁺ cells

(third column) that were CD19⁻ and CD20⁻ (lower gate, fourth column), CD16⁻ and CD14⁻ (fifth column), CD123⁻ (sixth column), CD4⁺ (seventh column), and CD32^{hi}, CD32^{int} or CD32^{lo} were then collected. Cells that were CD3⁺ and bearing markers of B cells (T-B; upper gate, fourth column) or other non-CD4-T-cells (T-other; combined ungated events from fifth and sixth columns) were also collected in separate tubes.

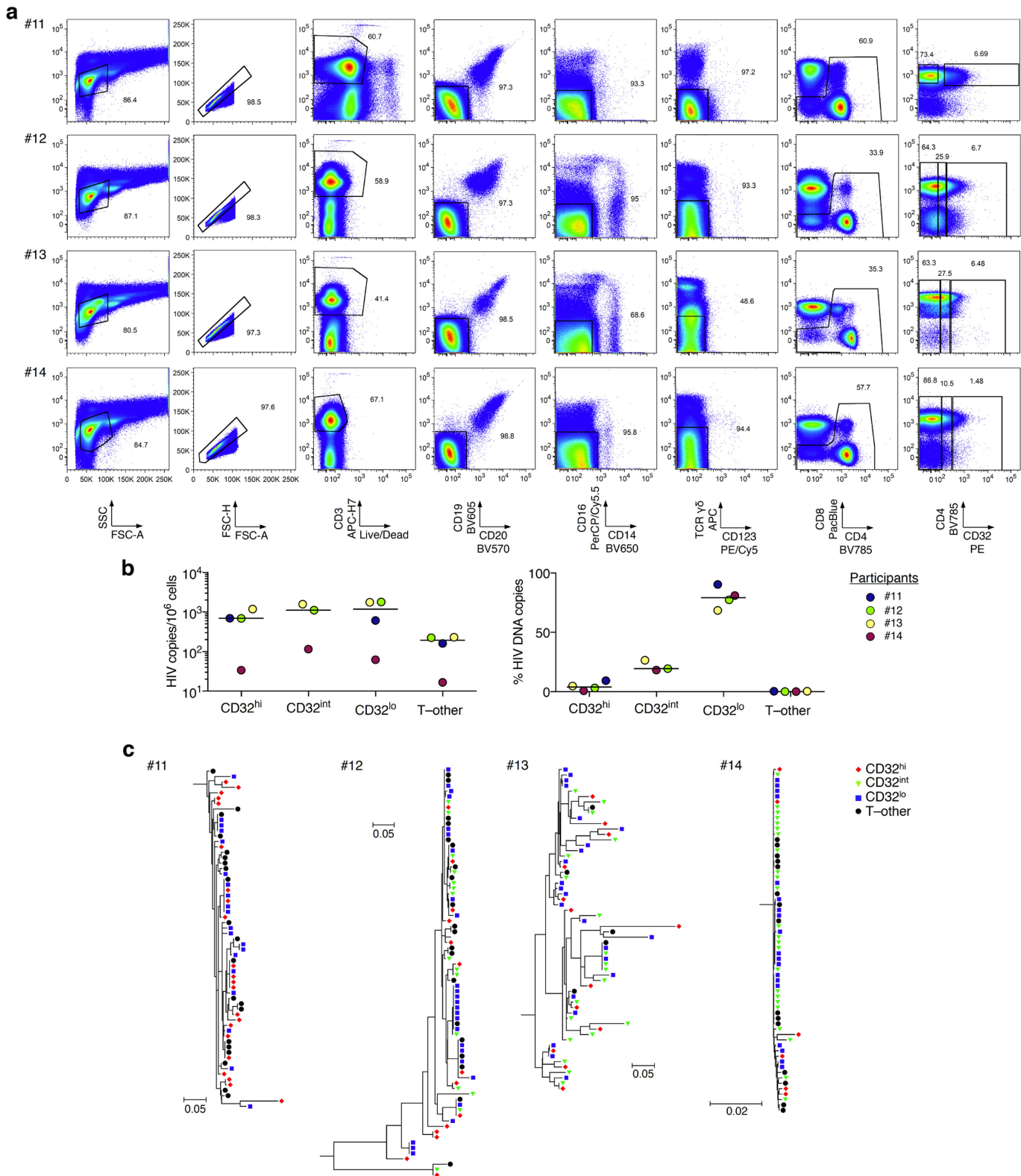
BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 4 | Post-sort flow cytometry of CD32 and CD4 expression by CD32^{hi}, CD32^{int}, CD32^{lo}, T-B and T-other cell subsets. Cells were sorted as in Extended Data Fig. 3. Note the large proportions of all CD32⁺ cells that did not show high CD4 expression after sorting.

Post-sort analyses of CD3⁺CD4⁺CD32^{hi} populations were deferred in cases in which these populations were too small to permit both post-sort analysis and downstream HIV DNA quantification (that is, donors # 1, 4 and 6–8).

BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 5 | See next page for caption.

BRIEF COMMUNICATIONS ARISING

Extended Data Fig. 5 | Flow cytometry, HIV DNA levels, and single-copy HIV DNA sequence analysis from CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations, and from T cells also bearing non-CD4-T-cell markers. **a**, PBMCs from four additional study participants were collected from whole blood by venipuncture with immediate processing (without cryopreservation). The T-other population was collected as a combination of the ungated events from CD19/CD20, CD16/CD14 and $\gamma\delta$ T cell receptor/CD123 exclusion plots (fourth, fifth and sixth columns). **b**, Left, copies of HIV DNA per million cells sorted from four additional study participants as in **a**. Right, percentages of all HIV DNA copies detected in

blood cells deriving from CD32^{hi}, CD32^{int}, CD32^{lo} and T-other subsets, calculated by adjusting values in the left panel for the relative proportions of these subsets determined using FACS data. **c**, Sequences of individual HIV DNA copies were determined by Sanger sequencing of products obtained by fluorescence-assisted clonal amplification, which amplifies a region of the HIV *env* gene. Phylogenetic trees were constructed as described in the Supplementary Methods. All Bonferroni-corrected Slatkin–Maddison *P* values for genetic compartmentalization between any two subsets were greater than 0.05 in all four participants.

The role of CD32 during HIV-1 infection

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

The persistence of latent HIV-1 in resting memory CD4⁺ T cells is a major barrier to a cure, and a biomarker for latently infected cells would be of great scientific and clinical importance^{1–5}. Using an elegant discovery-based approach, Descours et al.⁶ reported that CD32a, an Fcγ receptor not normally expressed on T cells, is a potential biomarker for the HIV-1 reservoir in CD4⁺ T cells⁶. Using a quantitative viral outgrowth assay (qVOA), we show that CD32⁺CD4⁺ T cells do not contain the majority of intact proviruses in the latent reservoir and that the enrichment found by Descours et al.⁶ may in part reflect the use of an ultrasensitive ELISA that does not predict exponential viral outgrowth. Our studies show that CD32 is not a biomarker for the major population of latently infected CD4⁺ T cells. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0496-1> (2018).

If CD32a is a biomarker for latent HIV-1 infection in CD4⁺ T cells, one that is never expressed on CD4⁺ T cells in the absence of HIV-1 infection, then a difference in the frequency of CD4⁺ T cells that express CD32 in HIV-1-infected individuals relative to the frequency in healthy donors is expected. We isolated CD4⁺ T cells from infected and uninfected donors by negative selection and analysed the expression of CD32 and CD4 by flow cytometry. In healthy donors, an average of 0.019% of CD4⁺ T cells was also CD32⁺ (Fig. 1a). This value is not significantly different from levels in HIV-1-infected individuals (Fig. 1a; average 0.011%, $P = 0.1143$) or from values previously reported by Descours et al.⁶ in HIV-1-infected individuals (0.016%, $P = 0.66$). Thus, CD32 does not seem to be a specific biomarker of latently infected CD4⁺ T cells.

To examine whether replication-competent proviruses were present in CD4⁺CD32^{hi} T cells, total CD4⁺ T cells were isolated by negative selection from six HIV-1⁺ individuals that were treated with suppressive anti-retroviral therapy (ART) for at least 6 months (Supplementary Table 1). Freshly isolated cells were stained and sorted to obtain CD4⁺CD32^{hi} and CD4⁺CD32[−] populations, which were analysed in qVOAs⁷ (Fig. 1b, protocol 1). The number of CD4⁺CD32^{hi} cells assayed for each subject is shown in Fig. 1c. On day 14, outgrowth was measured using a standard ELISA for the HIV-1 p24 antigen. CD4⁺CD32^{hi} wells from all subjects were negative for p24 on day 14, and remained negative after an additional week of culture. Conversely, outgrowth was observed in CD4⁺CD32[−] wells from all subjects on both days 14 and 21. The mean infected cell frequency, 1.37 infectious units per million cells (IUPM), was comparable to values previously measured in resting CD4⁺ T cells in several studies (0.03–3.00 IUPM in HIV-1-infected patients⁸, 0.97 IUPM in chronically infected patients⁹) and to values previously measured in the same subjects (mean value 1.33 IUPM) (Fig. 1d, Supplementary Table 2). If the enrichment of proviruses in CD32⁺ cells reported by Descours et al.⁶ was characteristic of replication-competent proviruses, then outgrowth from CD4⁺CD32^{hi} T cells should have been seen (Fig. 1e).

One possible explanation for the discrepancy between our results and those of Descours et al.⁶ is that some latent HIV-1 may be present in a previously undescribed population of CD4⁺ T cells that express CD32 together with other non-T-cell lineage markers. Such cells would be removed during the negative selection used to isolate CD4⁺ T cells. Therefore, we freshly isolated total CD4⁺ cells from infected donors on suppressive ART using two methods: negative selection to remove other lineages, leaving untouched CD4⁺ T cells, and positive selection for

cells expressing CD4 (Fig. 1b, protocol 2). Both CD4⁺ populations were analysed by qVOA. No significant differences were observed in the frequencies of latently infected cells (Fig. 1f). Furthermore, no significant differences in proviral DNA were observed between the purified cell populations (Fig. 1g). Because CD4 is required for HIV-1 entry into the host cell, cell populations obtained via positive selection for CD4 should include every latently infected CD4⁺ T cell. Given that neither the infected cell frequencies nor the levels of proviral DNA differed between the purified cell populations, we conclude that no additional sizable population of latently infected cells was recovered by positive CD4 selection.

In further studies, we used a cell sorting strategy identical to that of Descours et al.⁶ on samples freshly isolated from six subjects receiving ART treatment. Peripheral blood mononuclear cells (PBMCs) isolated from subjects were stained and sorted to obtain CD3⁺CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32[−] cell populations that were tested for latently infected cells by qVOA analysis. The numbers of CD3⁺CD4⁺CD32^{hi} cells assayed for each subject are shown in Fig. 1c and Supplementary Table 3. In addition, total CD4⁺ cells were obtained by staining PBMCs for CD4 and sorting for CD4⁺ cells (Fig. 1b, protocol 3). qVOA results showed that both the CD3⁺CD4⁺CD32[−] and the total CD4⁺ T cell populations had the same infected cell frequencies that were comparable to frequencies measured in other studies¹⁰. However, we observed no outgrowth in CD3⁺CD4⁺CD32^{hi} cultures (Fig. 1h, Supplementary Table 2).

We also analysed CD3⁺CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32[−] cells isolated by the method of Descours et al.⁶ for the presence of proviral DNA by qPCR. We found 89 copies of *gag* per million CD3⁺CD4⁺CD32[−] cells, which is similar to previous measurements in total CD4⁺ T cells¹¹. However, no proviral DNA was detected after DNA extraction from 39,000 CD3⁺CD4⁺CD32^{hi} cells and subsequent qPCR analysis (data not shown). This finding makes it highly unlikely that this cell population is enriched for HIV-1 to a level of more than one provirus copy per cell, as reported by Descours et al.⁶. We caution that the normalization of very low-level HIV-1 DNA measurements from qPCR reactions done with a low number of input cells could artificially produce apparent enrichments in HIV-1 DNA.

In a further attempt to explain the discordant qVOA results obtained in our studies and those of Descours et al.⁶, we tested whether the use of the ultra-sensitive p24 digital ELISA¹² and the low cell input can affect IUPM calculations, leading to erroneous overestimation of latent infection. qVOA culture supernatants were assayed for HIV-1 p24 using the ultrasensitive SIMOA p24 2.0 assay (Quanterix) on days 5, 9, 14 and 21. Using the lower limit of quantification (0.01 pg ml^{−1}) as the cut-off level, we found that two out of three qVOAs containing CD4⁺CD32^{hi} cells tested positive for p24 by this assay, even though the same wells were negative by standard ELISA, which is several orders of magnitude less sensitive (Fig. 2a). Exponential outgrowth is the hallmark of replication-competent viruses. In qVOA cultures of CD4⁺CD32[−] cells, only a fraction of the wells that were positive by SIMOA showed exponential outgrowth as determined by standard ELISA on day 21 (Fig. 2b). Importantly, CD4⁺CD32^{hi} culture wells that tested positive by SIMOA p24 assay showed no exponential outgrowth and had significantly lower levels of p24 (Fig. 2c). It is possible that low positive SIMOA values could reflect an assay artefact or the presence of defective proviruses that are still capable of producing low levels of Gag¹³. A further concern

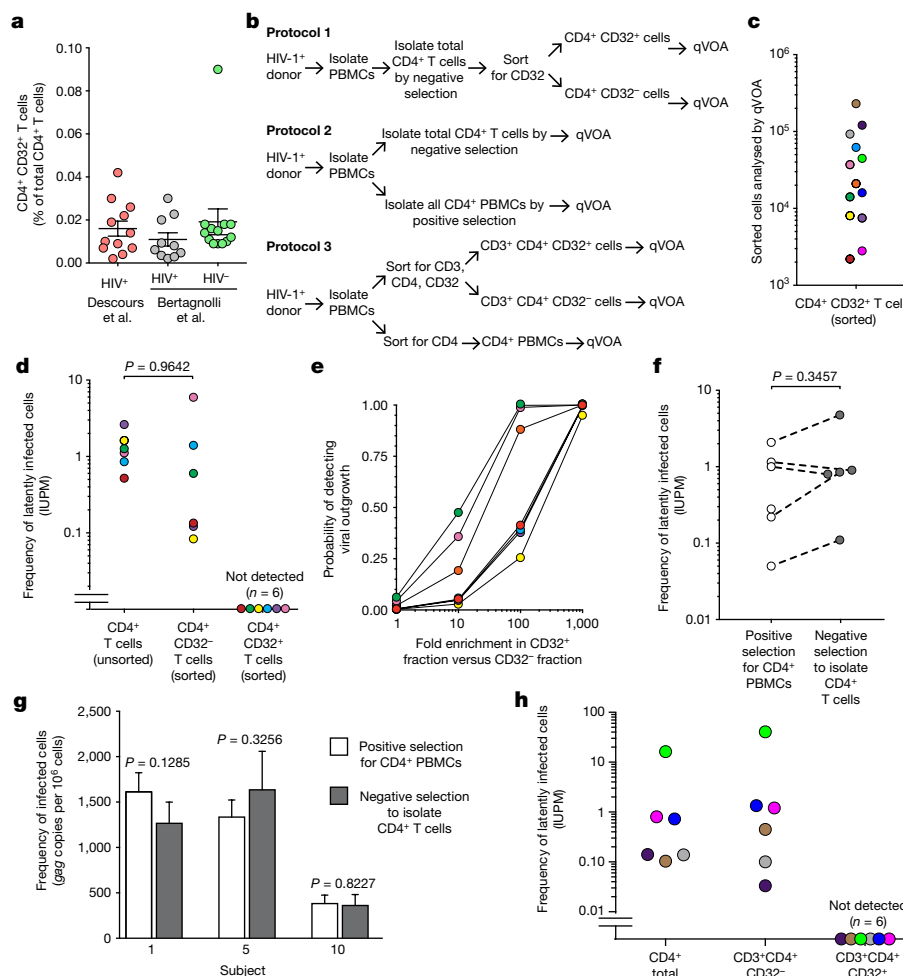


Fig. 1 | Analysis of CD4⁺CD32⁻ and CD4⁺CD32⁺ populations by qVOA and proviral DNA measurements. **a**, Percentage of CD4⁺CD32^{hi} T cells relative to total CD4⁺ T cells in healthy donors and HIV-1-infected donors. Infected donor values were obtained from supplementary table 4 of Descours et al.⁶. LLOQ, lower limit of quantification. **b**, Schematic depicting the three strategies (protocols 1–3) used to obtain different populations of CD4⁺ T cells analysed in qVOAs. **c**, Numbers of sorted CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32^{hi} T cells from each subject analysed in qVOAs. **d**, Frequencies of latently infected cells among CD4⁺CD32^{hi} T cells and CD4⁺CD32⁻ T cells and among total CD4⁺ T cells from the same subjects previously measured in separate experiments. Cells were isolated using protocol 1 (colours correspond to subject values from panel c). **e**, Probability of detecting outgrowth based on measured frequencies of latently infected cells among the CD4⁺CD32⁻ fraction and number of CD4⁺CD32^{hi} cells plated assuming various degrees of enrichment of HIV-1 in CD32^{hi} cells. **f**, Frequencies of latently infected cells measured in qVOAs using positive or negative selection to obtain total CD4⁺ cells (protocol 2; positive selection was accomplished by either sorting or CD4 microbead strategies, with similar results). **g**, Comparison of proviral DNA measurements obtained with qPCR on total CD4⁺ cells purified using positive or negative selection (protocol 2). **h**, Frequencies of latently infected cells among total CD4 cells, and CD3⁺CD4⁺CD32⁻ and CD3⁺CD4⁺CD32^{hi} populations. Cells were isolated using protocol 3 (colours correspond to subject values from panel c).

is that the IUPM calculations are based on cell input, fold dilutions and technical replicates¹⁴, and thus, qVOA analyses performed with very small numbers of sorted CD4⁺CD32^{hi} cells can markedly skew the frequency of cells harbouring replication-competent proviruses (five-fold dilutions from 800 to 1 cell in Descours et al.⁶). When we applied the results obtained with the SIMOA p24 assay, IUPM values ranged from 0 to 3,134 and 554 (patients 4 and 5, respectively; Fig. 2d). As a consequence, when we calculated the ‘fold enrichment’ of IUPM in the CD4⁺CD32^{hi} cells compared to the CD4⁺CD32⁻ cells, we observed a mean fold enrichment of 665 (range 152–1179, from the two patients with positive p24 using SIMOA), similar to what was reported by Descours et al.⁶ (Fig. 2e).

In summary, we find no evidence that CD32 expression indicates the presence of latent HIV-1, and demonstrate that at least a substantial fraction of the HIV-1 latent reservoir is in CD3⁺CD4⁺CD32⁻

T cells. Although no outgrowth could be found in cultures containing CD4⁺CD32^{hi} T cells, viral outgrowth comparable to historical measurements was found in cultures containing CD4⁺CD32⁻ T cells. The use of an ultrasensitive p24 ELISA assay may account for the apparent enrichment observed in culture experiments by Descours et al.⁶. In short, our results have demonstrated that CD32 does not define the HIV-1 reservoir and that future research is needed to identify biomarkers for latently infected cells.

We thank the study participants without whom this research would not be possible. Funding was provided by the US National Institutes of Health (NIH) Martin Delaney I4C, Beat-HIV and DARE Collaboratories by the Johns Hopkins Center for AIDS Research (P30AI094189), by NIH grant 43222, and by the Howard Hughes Medical Institute and the Bill and Melinda Gates Foundation.

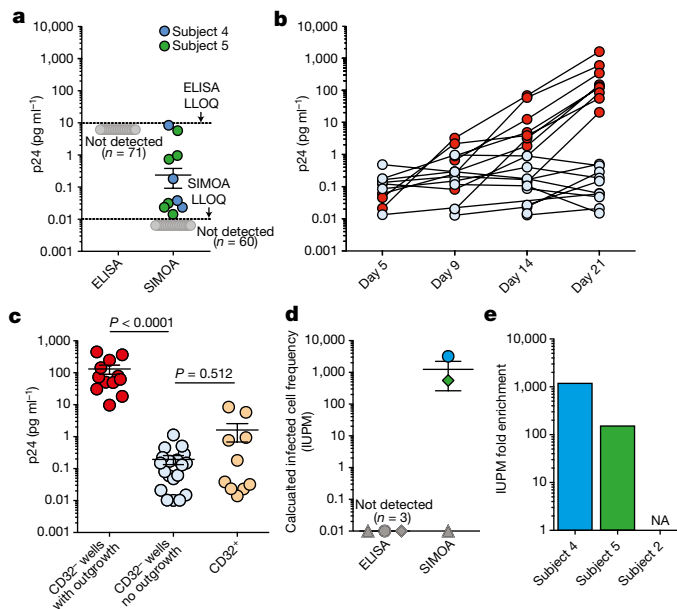


Fig. 2 | Ultrasensitive p24 measurements. **a**, Levels of p24 from CD32⁺ culture wells measured by ELISA and SIMOA (lower limit of quantification: 5–10 pg ml⁻¹ and 0.01 pg ml⁻¹, respectively) (data collected from three subjects, for a total of 71 wells). **b**, Longitudinal levels of p24 measured by SIMOA in individual culture wells in the qVOA for CD32⁻ cells from subject 5, showing wells with and without viral outgrowth (red and blue circles, respectively). **c**, Levels of p24 measured by ELISA in CD32⁻ wells with outgrowth compared with SIMOA measurements in wells with no outgrowth and CD32⁺ wells (data collected from subjects 2, 4 and 5). *P* values were determined with a non-parametric *t*-test. **d**, IUPM calculation based on ELISA and SIMOA analysis. Symbols in dark grey represent values below the limit of detection. **e**, Fold enrichment of IUPM in CD32⁺ cells (from subjects 2, 4 and 5). NA, not applicable.

Methods

qVOAs isolated CD4⁺ T cells using negative depletion and were sorted for CD32⁺ cells (Fig. 1b, protocol 1). To test whether negative depletion was causing a loss of CD32⁺ CD4⁺ T cells, outgrowth and proviral DNA were compared from qVOAs in which CD4⁺ T cells were isolated using positive selection to measurements using negative depletion. Outgrowth measurements and proviral DNA were also measured using the methods described by Descours et al.⁶. Proviral DNA measurements were performed using qPCR¹⁵. HIV-1 p24 values were measured using both a standard ELISA for p24 antigen (Perkin Elmer) and SIMOA (Quanterix). Further details are provided in Supplementary Methods.

Data availability. All data are available from the corresponding author upon reasonable request.

Lynn N. Bertagnoli¹, Jennifer A. White¹, Francesco R. Simonetti¹, Subul A. Beg^{1,2}, Jun Lai^{1,2}, Costin Tomescu³, Alexandra J. Murray¹, Annukka A. R. Antar¹, Hao Zhang⁴, Joseph B. Margolick⁴, Rebecca Hoh⁵, Stephen G. Deeks⁵, Pablo Tebas⁶, Luis J. Montaner³, Robert F. Siliciano^{1,2*}, Gregory M. Laird¹ & Janet D. Siliciano¹

¹Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ²Howard Hughes Medical Institute, Baltimore, MD, USA. ³The Wistar Institute, Philadelphia, PA, USA. ⁴Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. ⁵Division of HIV, Infectious Diseases and Global Medicine, University of California, San Francisco, CA, USA. ⁶Division of Infectious Diseases, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. *e-mail: rsiliciano@jhmi.edu

Received: 29 September 2017; Accepted: 3 April 2018;

Published online 19 September 2018.

1. Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
2. Chun, T. W. et al. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
3. Wong, J. K. et al. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
4. Richman, D. D. et al. The challenge of finding a cure for HIV infection. *Science* **323**, 1304–1307 (2009).
5. Deeks, S. G. et al. Towards an HIV cure: a global scientific strategy. *Nat. Rev. Immunol.* **12**, 607–614 (2012).
6. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
7. Laird, G. M., Rosenbloom, D. I., Lai, J., Siliciano, R. F. & Siliciano, J. D. Measuring the frequency of latent HIV-1 in resting CD4⁺ T cells using a limiting dilution coculture assay. *Methods Mol. Biol.* **1354**, 239–253 (2016).
8. Siliciano, J. D. et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4⁺ T cells. *Nat. Med.* **9**, 727–728 (2003).
9. Eriksson, S. et al. Comparative analysis of measures of viral reservoirs in HIV-1 eradication studies. *PLoS Pathog.* **9**, e1003174 (2013).
10. Crooks, A. M. et al. Precise quantitation of the latent HIV-1 reservoir: implications for eradication strategies. *J. Infect. Dis.* **212**, 1361–1365 (2015).
11. Besson, G. J. et al. HIV-1 DNA decay dynamics in blood during more than a decade of suppressive antiretroviral therapy. *Clin. Infect. Dis.* **59**, 1312–1321 (2014).
12. Passaes, C. P. & Sáez-Cirión, A. HIV cure research: advances and prospects. *Virology* **454–455**, 340–352 (2014).
13. Pollack, R. A. et al. Defective HIV-1 proviruses are expressed and can be recognized by cytotoxic T lymphocytes, which shape the proviral landscape. *Cell Host Microbe* **21**, 494–506.e4 (2017).
14. Rosenbloom, D. I. et al. Designing and interpreting limiting dilution assays: general principles and applications to the latent reservoir for human immunodeficiency virus-1. *Open Forum Infect. Dis.* **2**, ofv123 (2015).
15. Massanella, M., Gianella, S., Lada, S. M., Richman, D. D. & Strain, M. C. Quantification of total and 2-LTR (long terminal repeat) HIV DNA, HIV RNA and herpesvirus DNA in PBMCs. *Bio Protoc.* **5**, e1492 (2015).

Author contributions L.N.B., J.A.W., G.M.L., R.F.S. and J.D.S. designed experiments. S.A.B., C.T. and L.J.M. obtained samples. L.N.B., J.A.W., S.A.B., G.M.L., R.F.S., J.L., A.J.M., A.A.R.A. and J.D.S. performed experiments. R.F.S., H.J. and J.B.M. performed cell sorting. L.N.B., J.A.W., R.F.S., A.J.M., A.A.R.A., R.F.S. and J.D.S. analysed the data and wrote the manuscript.

Competing interests Declared none.

Additional information

Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.F.S.

<https://doi.org/10.1038/s41586-018-0494-3>

Evidence that CD32a does not mark the HIV-1 latent reservoir

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

A recent report by Descours et al.¹ suggests that the cell surface expression of the low affinity Fc receptor CD32a (also known as FcγRIIa) marks the replication-competent HIV-1 reservoir in CD4⁺ T cells from 12 HIV-1-infected participants receiving suppressive anti-retroviral therapy (ART)¹. We have undertaken considerable efforts to replicate these findings using peripheral blood mononuclear cells (PBMCs) from 20 HIV-1-infected, ART-suppressed participants (Extended Data Table 1). We found no evidence to suggest that CD32a marks a CD4⁺ T cell population enriched in either HIV-1 DNA or replication-competent HIV-1 in our study participants. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-s41586-018-0496-1> (2018).

To validate these findings, we adopted the same gating strategy as described by Descours et al.¹ to define CD4⁺ T cell populations (Supplementary Fig. 1a). The CD32 antigen was identified using the same antibody clone (FUN-2) as described by Descours et al.¹. We observed the same CD4⁺ T cell subsets that stained at a high cell surface density of CD32 (CD4⁺CD32^{high}), an intermediate cell surface density of CD32 (CD4⁺CD32^{int}), and a CD4⁺ T cell subset lacking CD32 expression (CD4⁺CD32^{neg}). We obtained frequencies of CD4⁺CD32^{high} T cells that ranged from 0.002% to 0.026%, with a median value (0.012%) that was identical to that reported by Descours et al.¹ (Extended Data Table 2 and Supplementary Fig. 1a). Notably, we confirmed that this same CD4⁺CD32^{high} population is also present in PBMCs isolated from eight healthy donors and exists at similar frequencies to that in HIV-1-infected samples ($P = 0.971$, Extended Data Fig. 1a).

Next, we assessed the amount of replication-competent HIV-1 isolated from the same 20 participants by measuring the infectious unit per million cells (IUPM) in CD4⁺ T cells (range 0.01–37.5, median 0.46). Participant CD4⁺CD32^{high} T cell populations were colour-coded in descending order, and then divided into quartiles that corresponded to the relative frequency of CD4⁺CD32^{high} cells present in these samples (Fig. 1a).

After cytometric sorting of the various CD4⁺CD32 subsets, we quantified HIV-1 DNA in each population (total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high}, Fig. 1b) using droplet digital PCR (ddPCR), as described in the Methods. We found no evidence of HIV-1 DNA enrichment in the CD4⁺CD32^{high} fraction. We observed no significant difference in HIV-1 DNA between any populations and the CD4⁺CD32^{high} T cell population ($P = 0.28$). In fact, levels of HIV-1 DNA in the CD4⁺CD32^{high} T cell subsets isolated from nine participants was at the assay limit of detection (Fig. 1b). After correction for cell input in the CD4⁺CD32^{high} fraction, as estimated DNA values, we saw no evidence for HIV-1 DNA enrichment (open symbols in Extended Data Fig. 1b).

We then compared the relative frequency of the CD4⁺CD32^{high} T cell populations and the viral replicative capacity (IUPM values) per participant, but no relationship between the two parameters was observed (Fig. 1c). All values have been tabulated in Extended Data Table 2.

The HIV-1 reservoir largely resides in quiescent CD4⁺ T cells^{2,3}. Therefore, we sought to confirm the activation status of the CD4⁺ T cell populations by measuring the frequency of the activation markers CD69, CD25 and HLA-DR on CD4⁺ T cell subsets from all

participants. We found that the CD4⁺CD32^{high} T cells were highly activated compared to the CD4⁺CD32^{neg} T cells ($P < 0.0001$). Notably, among the activation markers, HLA-DR was particularly enriched,

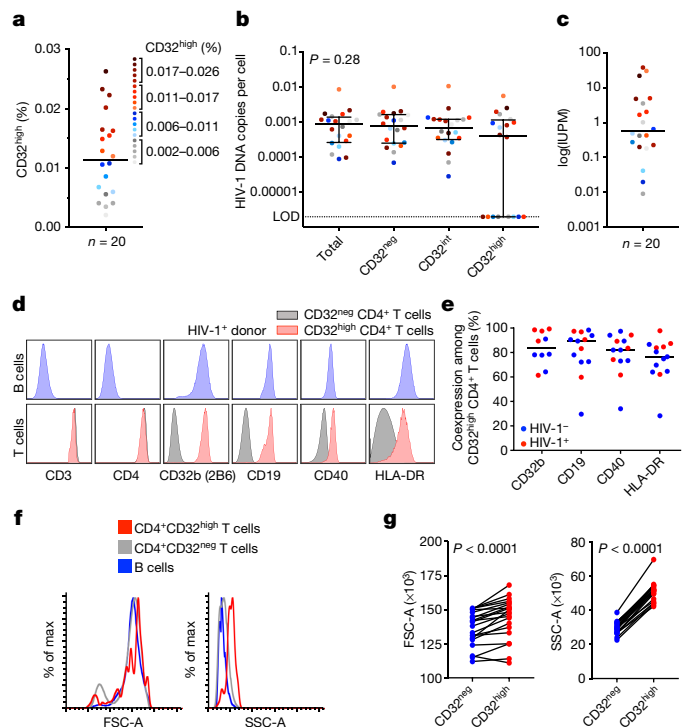


Fig. 1 | CD32-expressing CD4⁺ T cells are not enriched in HIV-1 DNA and express markers of B cell origin. **a–c**, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells from PBMCs of ART-suppressed, HIV-1-infected patients ($n = 20$) were sorted, and HIV-1 DNA was measured by ddPCR. **a**, Dividing the frequency (in percentage) of CD4⁺CD32^{high} T cells from all participants into quartiles, the values are shown as below or above the median. **b**, DNA copies per cell in sorted subsets of total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells are shown, with median and interquartile range (IQR). P value determined by Kruskal–Wallis test. LOD, limit of detection. **c**, IUPM in CD4⁺ T cells of each participant is shown in the colour corresponding to its frequency of CD4⁺CD32^{high} cells in panel **a**. **d**, **e**, CD32^{neg} and CD32^{high} (identified using FUN-2) CD4⁺ T cells from human PBMCs were assessed by flow cytometry for the expression of CD32b (2B6 antibody), CD19, CD40 and HLA-DR and compared to B cells (CD3⁺CD14⁺CD19⁺ lymphocytes). **d**, Representative flow cytometry results per cell antigen levels on B cells (top, blue histograms) and on CD32^{neg} and CD32^{high} CD4⁺ T cells (bottom, grey and red histograms, respectively) from PBMCs from an HIV-1⁺ participant. **e**, Frequency of CD4⁺CD32^{high} T cells staining positive for CD32b (2B6), CD19, CD40 or HLA-DR from HIV-1⁺ ($n = 5$) and HIV-1[−] ($n = 5–8$) human donor PBMC samples. Bars denote median values. **f**, Representative histograms of the FSC-A and SSC-A of B cells and CD32^{neg} and CD32^{high} CD4⁺ T cells from PBMCs of an HIV-1⁺, ART-suppressed participant sorted on a BD FACSAria II. **g**, Comparisons of the median FSC-A and SSC-A values between CD32^{neg} and CD32^{high} CD4⁺ T cell subsets from HIV-1⁺, ART-suppressed participants ($n = 20$). P values were determined using a paired t -test.

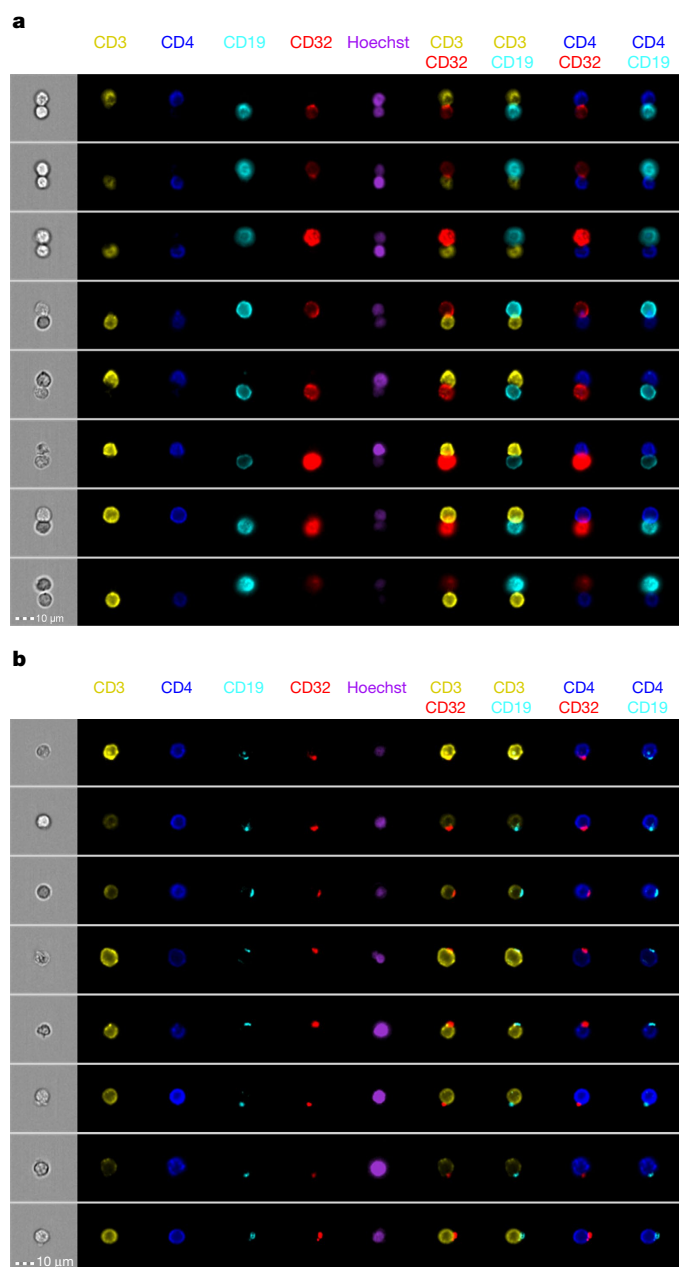


Fig. 2 | Flow cytometry imaging of sorted CD32-expressing T cells. **a**, Representative bright-field and pseudo-colour fluorescence images of T–B cell conjugates found in the CD32^{high} CD4⁺ T cell population sorted from PBMCs from HIV-1⁺, ART-suppressed participants, and imaged using Amnis technology. **b**, Representative images of punctate CD32 staining found on single T cells in the CD32^{high} and CD32^{int} population sorted from HIV-1⁺, ART-suppressed participant PBMCs.

marking approximately 75% of all CD4⁺CD32^{high} cells (median 74%), compared to CD4⁺CD32^{neg} cell populations (median 1.4%) (Extended Data Fig. 1c, $P < 0.0001$).

Two CD32 isoforms (CD32a and CD32b) are known to be expressed among all antigen presenting cells (APCs), but not typically on T cells. Therefore, we sought to exclude any APCs as potential contaminants of flow cytometry sorting. We evaluated the co-expression of lineage markers for all major CD32-bearing cells including monocytes, B cells, dendritic cells, granulocytes and natural killer cells. As expected, all CD32⁺ T cells expressed high amounts of CD3 and CD4 (Fig. 1d). However, we found that most CD4⁺CD32^{high} T cells from HIV-1⁺

patients, and also from healthy donors, co-expressed several B cell markers including CD19, CD40 and HLA-DR (Fig. 1d, e, Extended Data Fig. 2a). Notably, the B cell antigens found on CD4⁺CD32^{high} T cells were present at similar cell-surface densities as detected on bona fide B cells (Fig. 1d, e, Extended Data Fig. 2a).

The demonstration that the CD4⁺CD32^{high} fraction seen in HIV-1⁺ patients was marked with several B cell antigens and was similarly present in naive donors led us to investigate the origin of these B cell markers on a CD4⁺ T cell. Several reports have shown that B cells exclusively express the CD32b isoform⁴. The FUN-2 antibody clone used by Descours et al.¹ cannot distinguish between the CD32a and CD32b isoforms. Therefore, we used the monoclonal antibody clone 2B6 that has been reported to exclusively bind to CD32b^{4,5}. After co-staining PBMCs from HIV-1⁺ and HIV-1⁻ individuals with both the FUN-2 and the 2B6 antibodies, we found that all CD4⁺CD32^{high} T cells were marked only by the CD32b isoform and not by CD32a (Fig. 1d, e, Extended Data Fig. 2b), indicating that B cells are the origin of the CD32b antigen that marks the CD4⁺CD32^{high} T cells.

We sought to confirm this by determining whether *CD32A* (also known as *FCGR2A*) or *CD32B* (*FCGR2B*) mRNA was endogenously produced in the CD4⁺CD32^{high} subsets. After isolating total cellular RNA from various sorted T cell subsets, we used established reverse transcription PCR (RT–PCR) primers and probes that are specific to the CD32a and CD32b isoforms, as described in the Methods. We found that sorted CD4⁺CD32^{high} T cells from four HIV-1-infected participants did not contain detectable levels of the *CD32A* isoform. However, the *CD32B* mRNA isoform was readily detected in CD4⁺CD32^{high} T cells isolated from two out of four HIV-1⁺ patients (Extended Data Fig. 2c). By additional RT–PCR analysis, we detected both *CD3G* and *CD19* transcripts in the same CD4⁺CD32^{high} T population, indicating that the CD32b marking the CD4⁺CD32^{high} T cells may be from B cells expressing cognate CD32b (Extended Data Fig. 2d).

Because this may require cell-to-cell interaction, we performed a back-gating analysis of our flow cytometry data and confirmed that all CD4⁺CD32^{high} populations were identified within single-cell gates (Supplementary Fig. 1b). However, post-hoc analysis comparing the forward and side scatter light pulse area (FSC–A and SSC–A, respectively) values between CD4⁺CD32^{neg} and CD4⁺CD32^{high} T cells showed that the CD4⁺CD32^{high} populations had both a significantly higher FSC–A ($P < 0.0001$) and SSC–A ($P < 0.0001$), suggesting that the CD4⁺CD32^{high} population may consist largely of cell doublets (Fig. 1f, g).

We next used Amnis imaging flow cytometry to visualize the sorted CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} cell populations directly. As expected, the CD4⁺CD32^{neg} and the CD4⁺CD32^{int} cell populations each consisted of more than 99% single cells. However, the CD4⁺CD32^{high} fraction contained a high frequency of cell doublets (mean value 94%) (Extended Data Fig. 3). Of these ‘doublets’, approximately 70% seemed to be coincident doublets, and 30% were conjugates of T and B cells (Fig. 2a and Extended Data Fig. 3b).

We observed no examples in which CD32 staining on T cells was distributed throughout the cell membrane, supporting the idea that the CD32 found in the CD4⁺CD32^{high} population is not the result of endogenous expression from CD4⁺ T cells. Of the instances in which CD32 was detected on a T cell in the CD4⁺CD32^{high} population, the staining was punctate and often co-localized with punctate CD19 staining (Fig. 2b), suggesting that CD32 was acquired via contact between B and T cells. We noted that the frequency of T cells with punctate CD32 staining was substantially higher in the sorted CD32^{int} population. Thus, sorting for CD4⁺ T cells with a ‘high’ surface density of CD32 results in the selective enrichment of contaminating T–B cell doublets. As shown in Supplementary Fig. 1, these doublets cannot be discerned by routine cytometric FSC and SSC singlet gating strategies.

In summary, using samples from 20 HIV-1-infected, ART-suppressed participants, our data contradict the assertion that CD32a is a marker of

the replication-competent viral reservoir. Although we did detect similar frequencies of CD4⁺CD32^{high} populations to Descours et al.¹, we found no difference in the total HIV-1 DNA content between CD4⁺ T cell populations including or excluding the CD32^{high} fractions (Fig. 1b).

Notably, the CD4⁺CD32^{high} population was highly activated. Previous studies that have evaluated CD32 expression on T cells suggest that it may be detected after activation^{6,7} and led us to believe that this population may be atypical compared to a quiescent population harbouring the HIV-1 reservoir^{2,3}.

Our additional findings are incongruent with CD32a marking the replication-competent reservoir in CD4⁺ T cells; our phenotyping and RT-PCR experiments indicate that it is the CD32b isoform that marks the CD4⁺CD32^{high} cells (Fig. 1d, e, Extended Data Fig. 2b–d). This finding, combined with the demonstration that this cell population is found in uninfected individuals, conflicts with the assertion of Descours et al.¹ that CD32a is upregulated after the establishment of viral latency. Recent reports have corroborated the absence of CD32a transcripts in reactivated, clonal HIV-1-infected CD4⁺ T cells⁸.

The surface density of CD32b (and other B cell markers) on the CD4⁺CD32^{high} population was observed at similar densities to that on B cells. These data, combined with the post-hoc analysis, suggests that this population may be largely comprised of doublets. Direct interrogation of the CD4⁺CD32^{high} population via Amnis imaging confirmed that this population consisted largely of contaminating doublets; either co-incident events or cell-to-cell conjugates (Fig. 2a).

We demonstrate that the mechanism by which the CD32b isoform labels the CD4⁺CD32^{high} populations is through the direct interaction of CD4⁺ T and B cells, and possible trogocytotic transfer of B cell antigens to T cells, as observed in the CD4⁺CD32^{int} population (Fig. 2b). This may explain the transfer or membrane painting of antigens such as CD32b, CD40 and HLA-DR, among other markers^{9–11}. Not only have cell-to-cell membrane transfers been shown to occur commonly *in vivo* during viral infections, but such transfers largely occur on activated cells¹². Membrane-bound Fcγ receptors, including CD32b, are known to be extracted from APCs and then transferred to T cells, and serve as a surrogate of recent T cell and APC interactions¹³. Our demonstration of T–B cell conjugates in the CD4⁺CD32^{high} population and high levels of single cells in the CD4⁺CD32^{int} population support this notion (Fig. 2a, b).

Collectively, our findings confirm that selectively sorting for T cells with a high surface density of CD32 results in the enrichment of T–B cell doublet contaminants, which cannot be discerned by routine gating strategies. The true isoform, CD32b, that marks the CD4⁺CD32^{high} population is probably indicative of dynamic CD4⁺ T cell interaction with B cells, rather than a marker of the HIV-1 reservoir^{14,15}.

We thank S. Mordecai for Amnis technical expertise, and acknowledge support from NIAID grants AI091514, AI122942, AI127089 and AI131365 awarded to J.B.W. Support was also provided by the NIAID awarded Martin Delaney Collaboratory ‘BELIEVE’ grant AI126617, co-funded by NIDA, NIMH and NINDS awarded to D.F.N.

Methods

HIV-1⁺ participants were recruited through: The Maple Leaf Medical clinic in Toronto, Canada; The HIV Eradication and Latency (HEAL) cohort of Brigham and Women's and Massachusetts General Hospital; The Whitmann Walker Clinic in Washington, DC; or the Hospital of the University of Pennsylvania. The study was approved by the University of Toronto, The University of Pennsylvania and George Washington University ethics committees and according to the protocol approved by the Partners Human Research Committee and Institutional Review Board (IRB). Written informed consent was obtained from each participant.

The percentage of CD32⁺ (clone FUN-2) CD4⁺ T cells was measured in samples from study participants. Both CD32⁺ and CD32^{neg} CD4⁺ T cells were sorted and viral DNA was measured using ddPCR. The analysis of cell lineage markers by flow cytometry and RT-PCR was also conducted. Flow cytometry sorts from PBMCs used in HIV-1 DNA analyses were performed on cell subsets and assessed using Amnis imaging flow cytometry.

Data availability. All data and reagents are available from the corresponding author upon request.

Christa E. Osuna¹, So-Yon Lim¹, Jessica L. Kublin¹, Richard Apps², Elsa Chen¹, Talia M. Mota³, Szu-Han Huang³, Yanqin Ren³, Nathaniel D. Bachtel³, Athe M. Tsibris⁴, Margaret E. Ackerman⁵, R. Brad Jones³, Douglas F. Nixon³ & James B. Whitney^{1,6*}

¹Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. ²Center for Human Immunology, National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA. ³Division of Infectious Diseases, Weill Department of Medicine, Weill Cornell Medical College, New York, NY, USA. ⁴Brigham and Women's Hospital, Boston, Massachusetts Harvard Medical School, Boston, MA, USA. ⁵Thayer School of Engineering, Dartmouth College, Hanover, NH, USA. ⁶Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA. *e-mail: jwhitne2@bidmc.harvard.edu

Received: 11 August 2017; Accepted: 24 May 2018;
Published online 19 September 2018.

1. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
2. Chun, T. W. et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387**, 183–188 (1997).
3. Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
4. Veri, M. C. et al. Monoclonal antibodies capable of discriminating the human inhibitory Fcγ-receptor IIB (CD32B) from the activating Fcγ-receptor IIA (CD32A): biochemical, biological and functional characterization. *Immunology* **121**, 392–404 (2007).
5. Boruchov, A. M. et al. Activating and inhibitory IgG Fc receptors on human DCs mediate opposing functions. *J. Clin. Invest.* **115**, 2914–2923 (2005).
6. Engelhardt, W., Matzke, J. & Schmidt, R. E. Activation-dependent expression of low affinity IgG receptors FcγRII(CD32) and FcγRIII(CD16) in subpopulations of human T lymphocytes. *Immunobiology* **192**, 297–320 (1995).
7. Sandilands, G. P. et al. Differential expression of CD32 isoforms following alloactivation of human T cells. *Immunology* **91**, 204–211 (1997).
8. Cohn, L. B. et al. Clonal CD4⁺ T cells in the HIV-1 latent reservoir display a distinct gene profile upon reactivation. *Nat. Med.* **24**, 604–609 (2018).
9. Cone, R. E., Sprent, J. & Marchalonis, J. J. Antigen-binding specificity of isolated cell-surface immunoglobulin from thymus cells activated to histocompatibility antigens. *Proc. Natl Acad. Sci. USA* **69**, 2556–2560 (1972).
10. Hwang, I. et al. T cells can use either T cell receptor or CD28 receptors to absorb and internalize cell surface molecules derived from antigen-presenting cells. *J. Exp. Med.* **191**, 1137–1148 (2000).
11. Wetzel, S. A., McKeithan, T. W. & Parker, D. C. Peptide-specific intercellular transfer of MHC class II to CD4⁺ T cells directly from the immunological synapse upon cellular dissociation. *J. Immunol.* **174**, 80–89 (2005).
12. Rosenits, K., Keppler, S. J., Vucikuj, S. & Aichele, P. T cells acquire cell surface determinants of APC via *in vivo* trogocytosis during viral infections. *Eur. J. Immunol.* **40**, 3450–3457 (2010).
13. Daubeuf, S. et al. Preferential transfer of certain plasma membrane proteins onto T and B cells by trogocytosis. *PLoS One* **5**, e8716 (2010).
14. Garside, P. et al. Visualization of specific B and T lymphocyte interactions in the lymph node. *Science* **281**, 96–99 (1998).
15. Okada, T. et al. Antigen-engaged B cells undergo chemotaxis toward the T zone and form motile conjugates with helper T cells. *PLoS Biol.* **3**, e150 (2005).

Author contributions D.F.N. and J.B.W. designed the studies. R.B.J., R.A., E.C., Y.R., N.D.B., C.E.O., R.T. and S.Y.L. led the virology assays. S.H.H., D.C., J.L.K., M.A. and C.E.O. led the immunology assays. J.B.W. led the studies and wrote the paper with all co-authors.

Competing interests Declared none.

Additional information

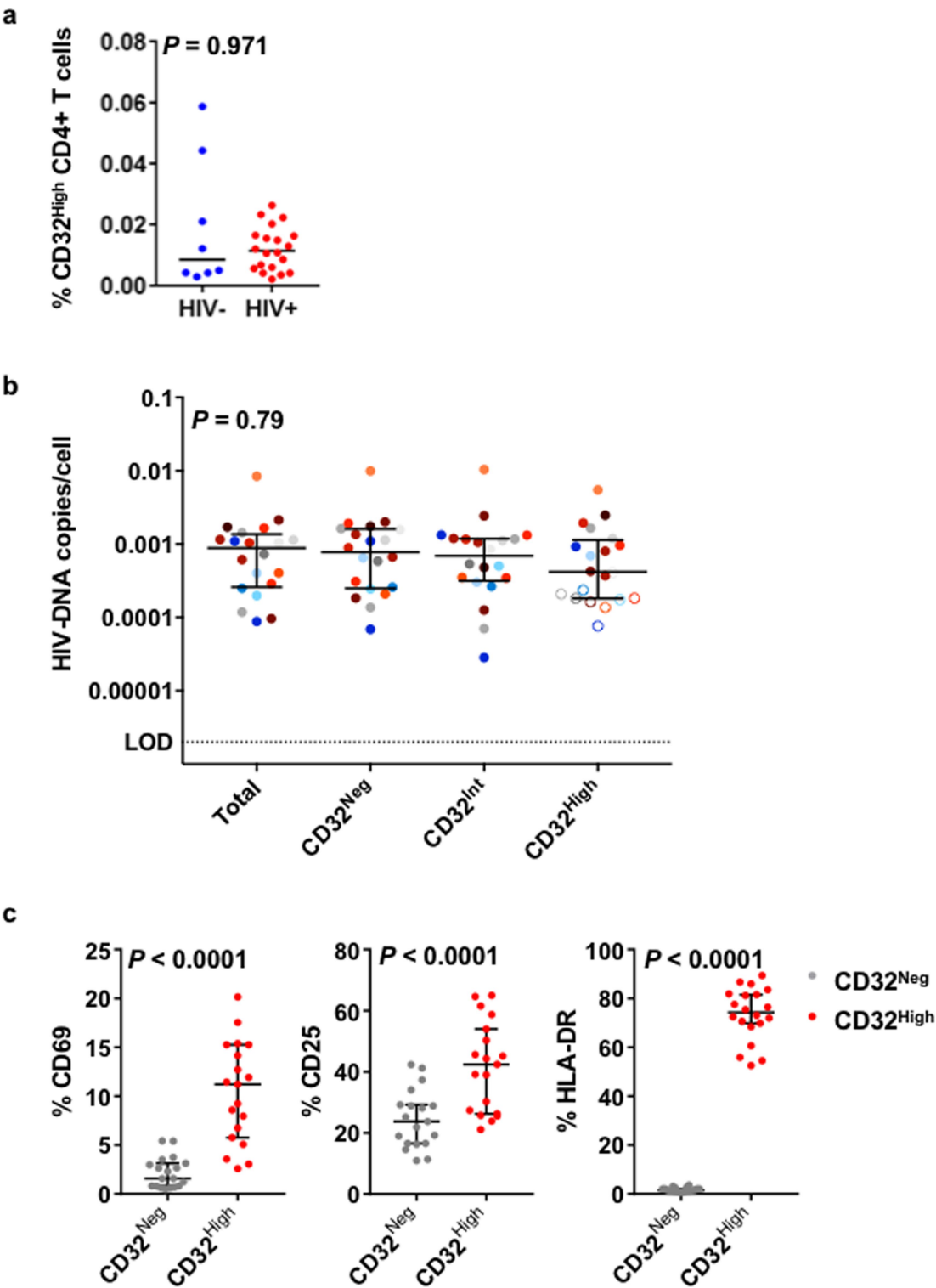
Extended data accompanies this Comment.

Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

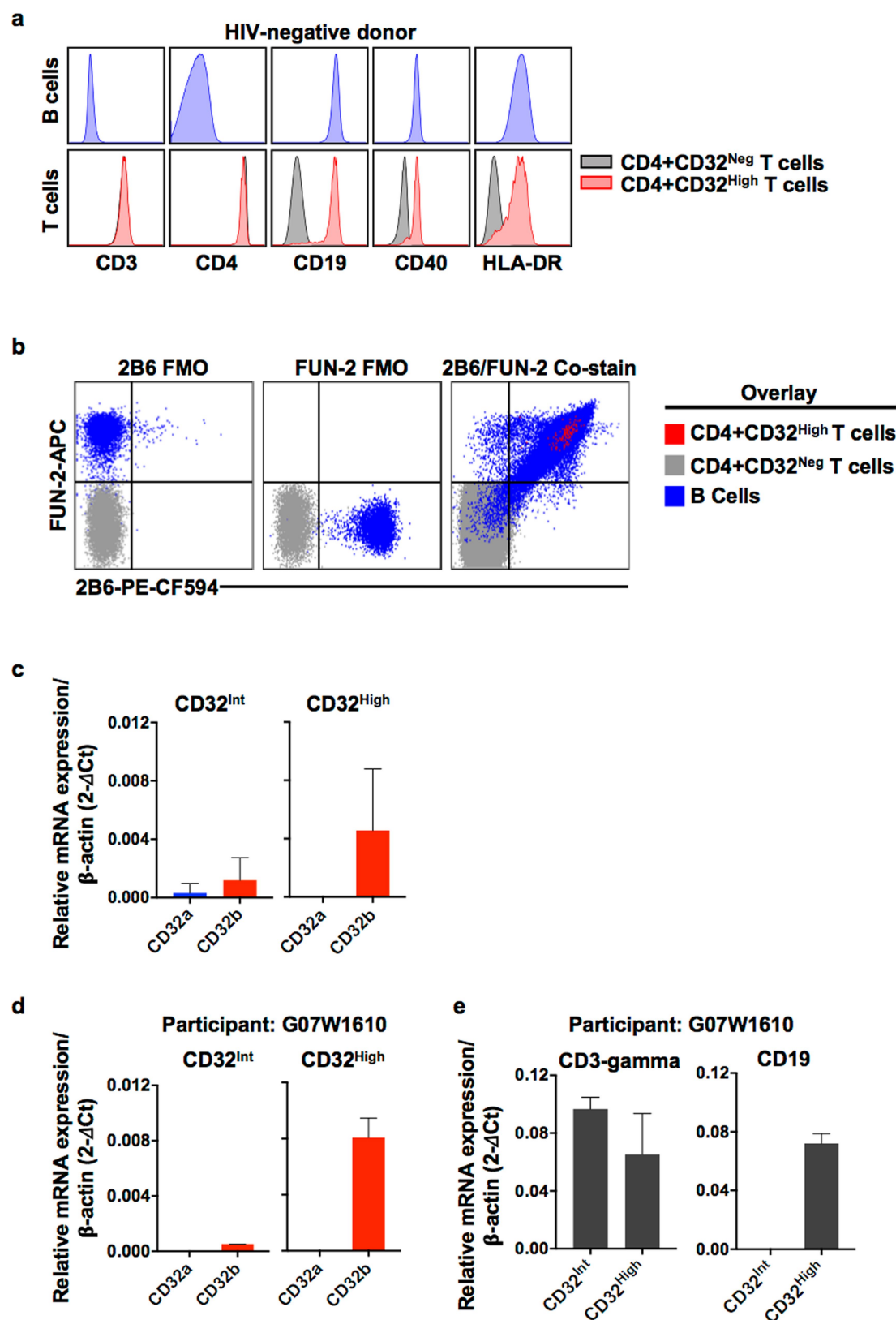
Correspondence and requests for materials should be addressed to J.B.W.

<https://doi.org/10.1038/s41586-018-0495-2>



Extended Data Fig. 1 | Frequency and activation status of CD32-expressing CD4⁺ T cells and their HIV-1 DNA content. **a**, The frequency of CD32^{high} CD4⁺ T cells was measured by flow cytometry in PBMCs from ART-suppressed, HIV-1⁺ ($n = 20$) and HIV-1⁻ ($n = 8$) donors. Bars denote median values. P values were determined by a Mann–Whitney test. **b**, DNA copies per cell in sorted subsets of total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells are shown with median values and the IQR. The results are shown as either the actual HIV-1 DNA

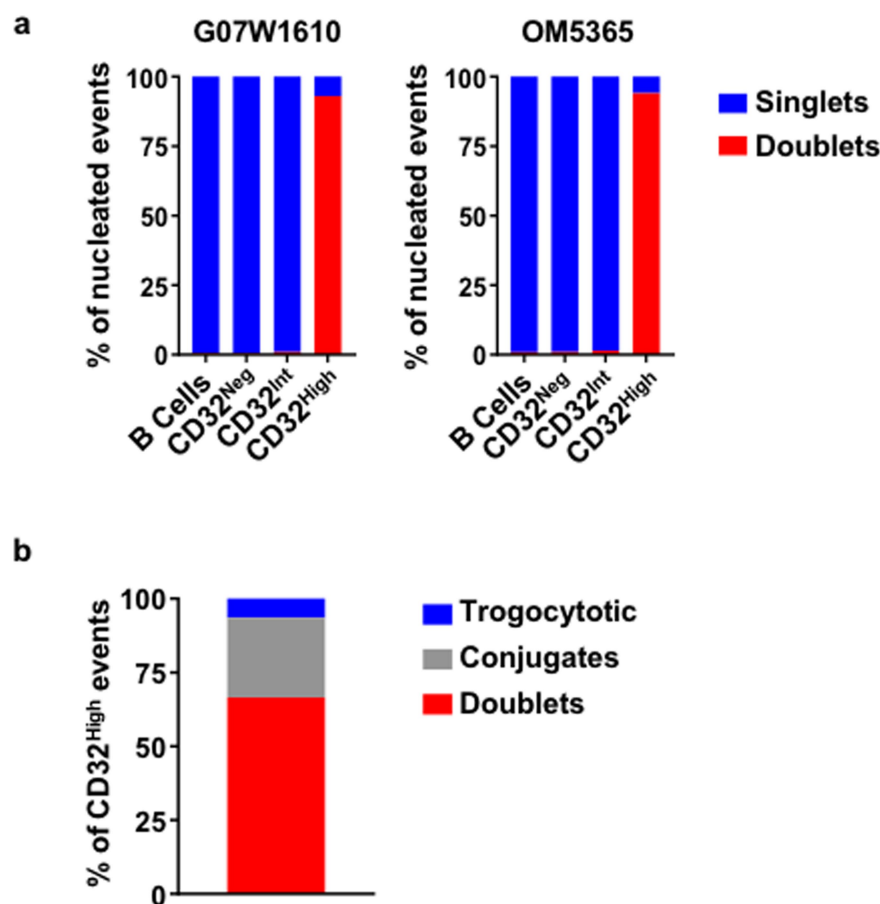
copies per million cells (filled symbols) or as estimated values calculated using the LOD and applied to the number of cells when the DNA input did not reach the threshold (open symbols). P values were determined by a Kruskal–Wallis test. **c**, The percentage of CD69, CD25 and HLA-DR expression was measured by flow cytometry on CD32^{neg} and CD32^{high} (FUN-2) CD4⁺ T cells from PBMCs from HIV-1⁺ participants ($n = 20$). Error bars show the median and IQR. P values were determined by Wilcoxon matched-pairs signed rank tests.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Detection of B cell proteins and mRNA in CD32-expressing CD4⁺ T cells. **a**, CD32^{neg} and CD32^{high} (FUN-2) CD4⁺ T cells from human PBMCs were assessed by flow cytometry for the expression of CD19, CD40 and HLA-DR, and compared to B cells (CD3⁻ CD14⁻ CD19⁺ lymphocytes). Representative flow cytometry results of per cell antigen levels on B cells (top, blue histograms) and CD32^{neg} and CD32^{high} CD4⁺ T cells (bottom, grey and red histograms, respectively) from an HIV-1⁻ donor. **b**, Representative CD32b staining of PBMCs from an HIV-1⁺, ART-suppressed participant. PBMCs were stained with an optimized concentration of the 2B6 monoclonal anti-CD32b antibody, followed by an antibody cocktail that included the FUN-2 monoclonal pan-CD32 antibody, as described in the Methods. Shown are the 2B6 and FUN-2

fluorescence minus one (FMO) antibody cocktail-stained samples and a sample co-stained with 2B6 and FUN-2. **c**, **d**, CD32 mRNA expression levels in CD4⁺CD32⁺ subsets. **c**, The relative expression of CD32A and CD32B mRNA isoforms in sorted CD4⁺CD32^{int} and CD4⁺CD32^{high} subsets from HIV-1⁺, ART-suppressed participants ($n = 4$). **d**, mRNA expression of CD32A and CD32B from patient G07W1610. **e**, T and B cell lineage-specific mRNA transcripts in sorted CD4⁺CD32⁺ subsets from participant G07W1610. Relative mRNA expression of target genes was normalized to *ATCB* using the comparative C_t method. Results are mean \pm s.d. of each value from each participant ($n = 4$; **c**), or from values generated from two separate experiments using samples from the same patient (**d**).



Extended Data Fig. 3 | Doublet composition of the sorted CD4⁺CD32^{high} T cells. Sorted B cells and CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells from an HIV-1⁺, ART-suppressed participant were analysed using an Amnis imaging cytometer. Singlets and doublets were quantified using the aspect ratio and nuclear staining. **a**, The proportion of total singlet and doublet events among total nucleated

cells detected on the Amnis cytometer in each sorted population was determined, and is shown as individual composite bar graphs for two patients (G07W1610 and OM5365). **b**, A composite bar graph of the proportion of conjugates, doublets and trogocytotic events that comprised the sorted CD4⁺CD32^{high} population ($n = 2$).

BRIEF COMMUNICATIONS ARISING

Extended Data Table 1 | Viral suppression of 20 HIV-1-infected participants on ART

Cohort	Participant ID	Date of Initial Suppression (MM/YY)	Length of suppression (yrs)
HEAL	HEAL-009	3/14	3
	HEAL-019	8/09	8
	HEAL-020	3/08	9.3
	HEAL-034	11/05	11.5
	HEAL-053	11/16	1
	HEAL-055	11/00	17
Maple Leaf	CIRC0024	6/98	17.0
	CIRC0133	7/08	7.0
	CIRC0196	4/14	1.2
	OM5011	11/08	6.6
	OM5148	1/08	7.5
	OM5162	9/04	10.8
	OM5203	3/12	3.3
	OM5334	7/14	0.9
	OM5365	3/08	7.3
WWH	WWH-B001	7/11	6.4
	WWH-B005	12/17	0.3
	WWH-B008	11/14	3.1
	WWH-B011	11/11	6
UPenn	G07W1610	10/05	11.8

BRIEF COMMUNICATIONS ARISING

Extended Data Table 2 | CD4⁺CD32^{high} subset proportions and HIV-1 DNA compared to total CD4⁺ and CD32^{neg} CD4⁺ T cells

Participant ID	CD32 ^{High}		HIV-DNA enrichment			
	% in total CD4	Absolute cell count	HIV-DNA copies/cell ¹	CD32 ^{High} /CD4 total ²	CD32 ^{High} /CD32 ^{Neg2}	CD32 ^{Neg} /CD4 total
HEAL-009	0.007	11,427	>0.000002	0.010	0.008	1.236
HEAL-019	0.002	4,911	>0.000002	0.002	0.001	1.500
HEAL-020	0.011	8,482	>0.000002	0.008	0.008	1.036
HEAL-034	0.022	8,238	0.000426	0.199	0.212	0.937
HEAL-053	0.004	9,544	>0.000002	0.017	0.015	1.161
HEAL-055	0.011	21,806	0.00037	0.604	0.555	1.088
CIRC0024	0.015	26,200	>0.000002	0.023	0.029	0.782
CIRC0133	0.017	14,602	>0.000002	0.005	0.010	0.517
CIRC0196	0.006	5,935	0.000694	1.722	1.066	1.615
OM5011	0.008	8,862	0.001942	1.871	2.187	0.855
OM5148	0.004	8,788	0.001191	1.043	1.049	0.994
OM5162	0.016	7,133	0.000923	0.842	0.842	1.000
OM5203	0.026	12,254	>0.000002	0.021	0.011	1.911
OM5334	0.016	6,275	0.000959	0.579	0.499	1.160
OM5365	0.006	11,027	>0.000002	0.003	0.003	0.803
WWH-B001	0.020	10,922	>0.000002	0.007	0.006	1.076
WWH-B005	0.013	5,964	0.00547	0.650	0.550	1.182
WWH-B008	0.023	6,464	0.000805	0.696	0.595	1.169
WWH-B011	0.004	5,953	0.001653	1.154	1.016	1.135
G07W1610	0.012	7,984	0.002482	1.452	1.411	1.029
Median	0.012	8,635	0.000398	0.389	0.356	1.082

¹Values below the LOD (2 copies per 10⁶ cells) are shaded in grey.

²To calculate HIV-1 enrichment, 0.000002 was used for all values below the LOD.

Descours et al. reply

REPLYING TO L. Pérez et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0493-4> (2018); C. E. Osuna et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0495-2> (2018); L. N. Bertagnolli et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0494-3> (2018)

In our previous work¹, we used an in vitro model of HIV-infected unstimulated CD4 T cells to identify CD32 as a candidate marker of HIV⁺ resting CD4 T cells in vitro, and a subset of HIV⁺ total CD4 T cells containing replication-competent viruses in individuals that underwent anti-retroviral therapy (ART). Of note, we did not explore the transcriptional status of hosted viruses (latent or active) ex vivo, nor the activation state of these cells (quiescent or activated)¹. In the accompanying Comments^{2–4}, colleagues attempted to reproduce these findings. They present experiments that support the following conclusions: (1) the isolation of the CD32⁺ CD4 T cell population results from artefacts caused by the flow cytometry sorting method^{2,3}, and (2) the sorted CD32 CD4 T cell population is not enriched in HIV nor in replication-competent proviral DNA^{2–4}. Here, we formulate two questions that mirror the major issues raised by these three Comments^{2–4} and discuss their results in the context of our previous report¹ and more recently published studies.

Is there any evidence that a CD4 T cell can express CD32 in the context of HIV infection? This question is raised by both Osuna et al.² and Pérez et al.³. A recent report⁷, using in situ hybridization (which avoids the criticism of artefacts caused by flow cytometry sorting), showed that HIV-1 RNA co-localized with CD32A (also known as FCGR2A) RNA in 90% of examined cells in B cell follicles from four individuals. Because HIV primarily targets CD4 T cells, these data may support the ability of a CD4 T cell to upregulate CD32 mRNA transcription after infection in vivo. Three independent groups have identified CD32 as being expressed by latently or productively infected CD4 T cells in vitro^{1,5–7}. These models generated and analysed a substantial percentage of HIV-infected CD4 cells. Thus, any marker that is usually not expressed by CD4 T cells but that is detected at the surface of these cells after infection is unlikely to result from biased analyses of cellular doublets, as could be the case when working on rare events from ex vivo samples^{2,3}. Instead, these data suggest that transcriptional regulation leading to the expression of CD32 mRNA and protein can probably occur after in vitro and in vivo infection of a single CD4 T cell.

Does the CD32 CD4 T cell subset contribute to viral persistence under treatment? All three of the accompanying Comments^{2–4} indicate that CD32 CD4 T cells are not enriched for HIV DNA in blood. Recent work suggests, however, that in some virally suppressed HIV-infected individuals, CD32 CD4 T cells were enriched in HIV DNA, although to a lesser extent than we reported⁸. Notably, this question has been recently addressed in tissues, and results seem to be less contrasted than in blood^{7,9,10}. More importantly, they revealed functional properties of these reservoir cells that have not been previously explored^{7,9,10}. As discussed above, a recent report⁷ found that within the B cell follicles of virally suppressed HIV-infected individuals, most of the cells containing HIV RNA and persisting despite treatment were found to express CD32A RNA⁷. This result seems to be in line with other data¹⁰ that indicate that T follicular helper cells, primarily found in these territories, were enriched for HIV DNA and RNA when expressing CD32¹⁰, although at a lower extent than our previous findings¹. In non-lymphoid rectal tissue, CD4 T cells expressing CD32 were also enriched

for both HIV DNA and RNA⁹. Notably, the co-expression of CD32 and HIV RNA reported in these two publications^{9,10} suggests that CD32 marks transcriptionally active infected cells rather than latent cells. Together, these reports support the ability of CD32 to identify a subset of persistent HIV-infected CD4 T cells and suggest that they could contribute to viral persistence under ART in vivo.

In conclusion, we believe that rather than completely ruling out the relevance of CD32 for the identification of a subset of infected cells in vivo and their contribution to HIV persistence, the whole literature, including the three accompanying Comments^{2–4}, opens new technical challenges and questions that we should solve in the near future.

Benjamin Descours, Gael Petitjean and Monsef Benkirane are solely responsible for this Reply. The contributions of the remaining authors from the original Letter¹ were limited to recruiting patients or performing analysis on blinded samples, and thus only Descours, Petitjean and Benkirane have authored this Reply.

Benjamin Descours¹, Gael Petitjean¹ & Monsef Benkirane^{1*}

¹Institut de Génétique Humaine, Laboratoire de Virologie Moléculaire, UMR9002, CNRS, Université de Montpellier, Montpellier, France.

*e-mail: monsef.benkirane@igh.cnrs.fr

1. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
2. Osuna, C. E. et al. Evidence that CD32a does not mark the HIV-1 latent reservoir. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0495-2> (2018).
3. Pérez, L. et al. Conflicting evidence for HIV enrichment in CD32⁺ CD4 T cells. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0493-4> (2018).
4. Bertagnolli, L. N. The role of CD32 during HIV-1 infection. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0494-3> (2018).
5. Iglesias-Ussel, M., Vandergeeten, C., Marchionni, L., Chomont, N. & Romero, F. High levels of CD2 expression identify HIV-1 latently infected resting memory CD4⁺ T cells in virally suppressed subjects. *J. Virol.* **87**, 9148–9158 (2013).
6. Grau-Expósito, J. et al. A Novel single-cell FISH-flow assay identifies effector memory CD4⁺ T cells as a major niche for HIV-1 transcription in HIV-infected patients. *MBio* **8**, e00876-17 (2017).
7. Abdel-Mohsen, M. et al. CD32 is expressed on cells with transcriptionally active HIV but does not enrich for HIV DNA in resting T cells. *Sci. Transl. Med.* **10**, eaar6759 (2018).
8. Martin, G. E. et al. CD32-expressing CD4 T cells are phenotypically diverse and can contain proviral HIV DNA. *Front. Immunol.* **9**, 928 (2018).
9. Hogan, L. E. et al. Increased HIV- transcriptional activity and infectious burden in peripheral blood and gut-associated CD4⁺ T cells expressing CD30. *PLoS Pathog.* **4**, e006856 (2018).
10. Noto, A., Procopio, F., Corpataux, J. M. & Pantaleo, G. CD32⁺PD1⁺ Tfh cells are the major HIV reservoir in long-term art-treated individuals. *J. Virol.* <https://doi.org/10.1128/JVI.00901-18> (2018).

Author contributions B.D., G.P. and M.B. wrote the manuscript.

Competing interests Declared none.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.B.

<https://doi.org/10.1038/s41586-018-0496-1>

Conflicting evidence for HIV enrichment in CD32⁺ CD4 T cells

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

Descours and colleagues¹ reported a marked enrichment for HIV among CD32a⁺ CD4 T cells in people receiving anti-retroviral therapy (ART). This tiny CD32a⁺ population (0.012% of all blood CD4 T cells) contained a median of 0.56 HIV DNA genomes per cell, and accounted for 26.8–86.3% of HIV DNA in CD4 T cells, thus suggesting that targeting CD32a⁺ CD4 T cells might help to clear HIV reservoirs in vivo. Here, we report our unsuccessful attempts to confirm these findings. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0496-1> (2018).

We first used fluorescence-activated cell sorting (FACS) to sort CD4 T cells with high, intermediate and low levels of CD32 staining (CD32^{hi}, CD32^{int} and CD32^{lo}, respectively) from 10 individuals with chronic HIV infection who were receiving ART (mean duration, 8.8 years; range, 2.7–15). We used cell-staining reagents and gating techniques that matched those used by Descours et al.¹ (see Supplementary Methods and Extended Data Fig. 1). As shown in Fig. 1a, we detected no enrichment for HIV DNA in the CD32^{hi} or CD32^{int} CD4 T cells. Moreover, the CD32^{hi} and CD32^{int} subsets combined accounted for no more than 3% of all HIV DNA copies within circulating CD4 T cells in any of the 10 study participants (Fig. 1b). Post-sort flow cytometry of CD32^{hi} and CD32^{int} populations showed heterogeneous patterns that suggested the formation of T cell–B cell or T cell–monocyte conjugates as the origin of most CD32^{hi} or CD32^{int} CD4 T cells, with separation of these conjugates during sorting (Extended Data Fig. 2).

To rule out the possibility that we had inadvertently obtained false negative results either by excluding HIV-infected, CD32⁺ CD4 T cells using tight light scatter gates or by failing to exclude non-T-cell contaminants, we performed parallel sorts on the same 10 samples using an alternative gating scheme. We used a more inclusive light scatter gate as well as markers for B cells, monocytes, dendritic cells and natural killer cells (Extended Data Fig. 3). Events that were CD3⁺ were separated into fractions that were positive for B cell markers (T–B), positive for one or more other non-CD4-T-cell markers (T–other), or negative for all of these, positive for CD4, and CD32^{hi}, CD32^{int} or CD32^{lo}. Neither CD32^{hi} nor CD32^{int} CD4 T cells were enriched for HIV DNA (Fig. 2a). Similarly, we detected no enrichment for HIV DNA in the T–B and

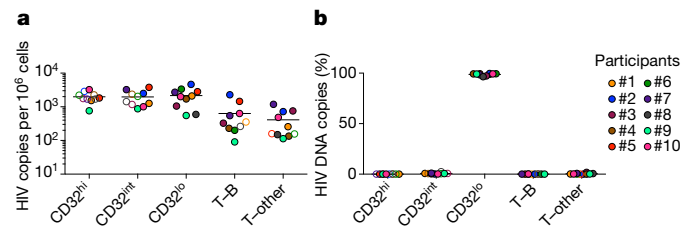


Fig. 2 | Levels of HIV DNA in CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cells, sorted using alternative gating. The samples from Fig. 1 were sorted using alternative gating in which T cells bearing markers of B cells (T–B) or other non-CD4-T-cell lineages (T–other) were first collected in separate tubes. **a**, Copies of HIV DNA per million sorted cells. **b**, Percentages of all HIV DNA copies detected in blood cells that were detected within each subset, calculated by adjusting values in **a** for the relative proportions of these subsets in FACS data.

T–other populations (Fig. 2a). In each of the 10 participants, at least 96% of all HIV DNA copies occurred in conventional CD32^{lo} cells (Fig. 2b). Post-sort flow cytometry suggested that most events bearing both T-cell and non-CD4-T-cell markers again represented cell–cell conjugates, and also showed that most remaining CD32^{hi} CD4 T cells did not reproducibly show a high CD32 signal after sorting (Extended Data Fig. 4). This was in contrast to conventional CD32^{lo} cells, which were uniformly pure in post-sort analyses across participants. In a second group of four individuals whose peripheral blood mononuclear cells (PBMCs) were sorted without previous cryopreservation (Extended Data Fig. 5a), we again found no enrichment for HIV DNA based on CD32 expression (Extended Data Fig. 5b), and also observed that HIV DNA sequences in CD32⁺ CD4 T cells were genetically intermingled with HIV DNA sequences in other CD4 T cells (Extended Data Fig. 5c).

Overall, our studies showed no enrichment for HIV DNA in CD32⁺ CD4 T cells, and also raised questions about the source of the CD32 labelling on these cells. We propose that the CD32 expression associated previously with CD4 T cells could have arisen from adherent non-T-cells or cellular material bearing this marker, and that conjugates containing HIV-infected CD4 T cells could be differentially produced and/or recovered in different laboratories with different sample processing and FACS practices. It is important to acknowledge that these considerations do not explain the discrepancy between the Descours et al. study¹ and ours in the quantities of HIV DNA detected within CD3⁺CD4⁺CD32⁺ sorted material. Nevertheless, we wish to emphasize that our findings do not support targeting CD32 molecules on CD4 T cells in emerging HIV cure strategies.

Methods

Participant recruitment and informed consent were performed under Institutional Review Board (IRB)-approved protocols at the US National Institutes of Health (NIH). For FACS, whole PBMCs were stained with monoclonal antibodies matching those used by Descours et al.¹ (see Supplementary Methods) and sorted on a BD FACSARIA. To evaluate purity, a portion of each population was re-analysed on the flow cytometer after sorting. Virus DNA copies in sorted cells were enumerated by fluorescence-assisted clonal amplification². DNA recovery was quantified by albumin (*ALB*) quantitative PCR. Because the FUN-2 monoclonal antibody used

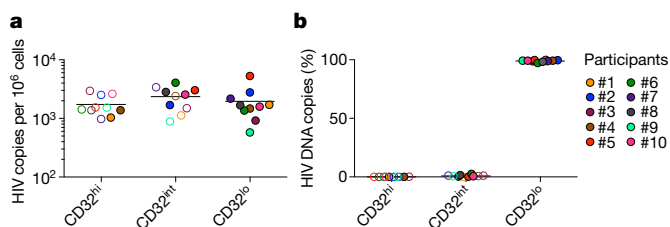


Fig. 1 | Levels of HIV DNA in CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cells, sorted from PBMCs of 10 ART-treated participants, as in Descours et al.¹ **a**, Copies of HIV DNA per million sorted cells. **b**, Percentages of all HIV DNA copies detected in blood CD4 T cells that were detected within each subset, calculated by adjusting values in **a** for the relative proportions of these subsets in FACS data. In all figures, horizontal bars denote median values, and open symbols indicate detection limits for measurements in which HIV DNA was not detected.

BRIEF COMMUNICATIONS ARISING

by Descours et al.¹ and in our study may recognize both CD32a and CD32b, we refer to cells staining with this monoclonal antibody as CD32⁺.

Data availability. All DNA sequences in this manuscript (analysed in Extended Data Fig. 5) have been deposited in GenBank under accession numbers MH080310–MH080572.

Liliana Pérez¹, Jodi Anderson², Jeffrey Chipman³, Ann Thorkelson², Tae-Wook Chun⁴, Susan Moir⁴, Ashley T. Haase⁵, Daniel C. Douek⁶, Timothy W. Schacker^{2,7} & Eli A. Boritz^{1,7*}

¹Virus Persistence and Dynamics Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. ²Division of Infectious Diseases, University of Minnesota, Minneapolis, MN, USA. ³Department of Surgery, University of Minnesota, Minneapolis, MN, USA. ⁴Laboratory of Immunoregulation, National Institute of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD, USA. ⁵Department of Microbiology and Immunology, University of Minnesota, Minneapolis, MN, USA. ⁶Human Immunology Section, Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. ⁷These authors jointly supervised this work: Timothy W. Schacker, Eli A. Boritz. *e-mail: boritze@mail.nih.gov

Received: 11 October 2017; Accepted: 20 March 2018;

Published online 19 September 2018.

1. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
2. Boritz, E. A. et al. Multiple origins of virus persistence during natural control of HIV infection. *Cell* **166**, 1004–1015 (2016).

Author contributions Data generation and analysis: L.P., J.A., T.W.S. and E.A.B. Study design and oversight: L.P., A.T.H., D.C.D., T.W.S. and E.A.B. Participant cohort and sample management: J.A., J.C., A.T., T.W.C., S.M. and T.W.S. Manuscript preparation: L.P., A.T.H., D.C.D., T.W.S. and E.A.B.

Competing interests Declared none.

Additional information

Extended data accompanies this Comment.

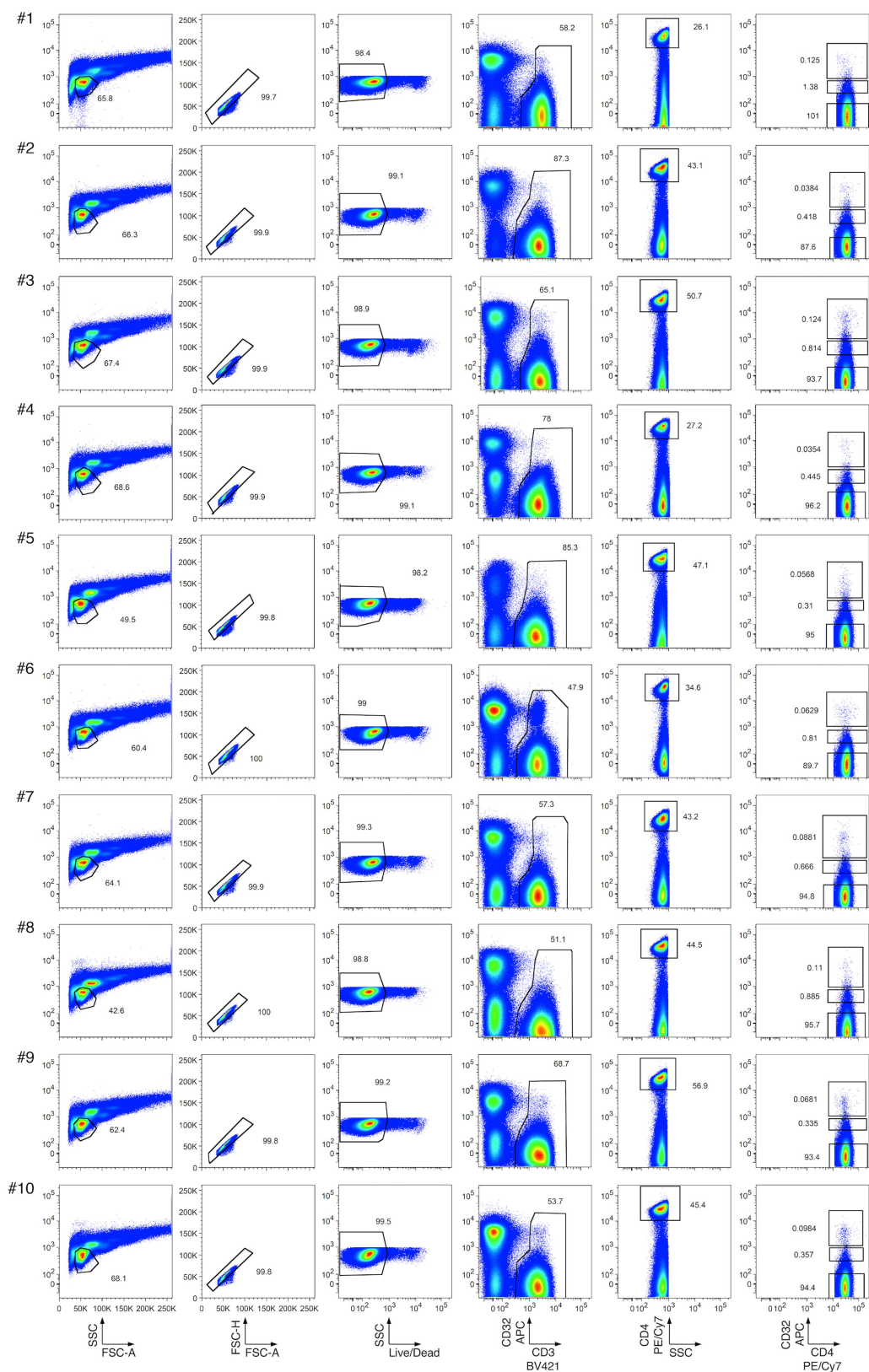
Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to E.A.B.

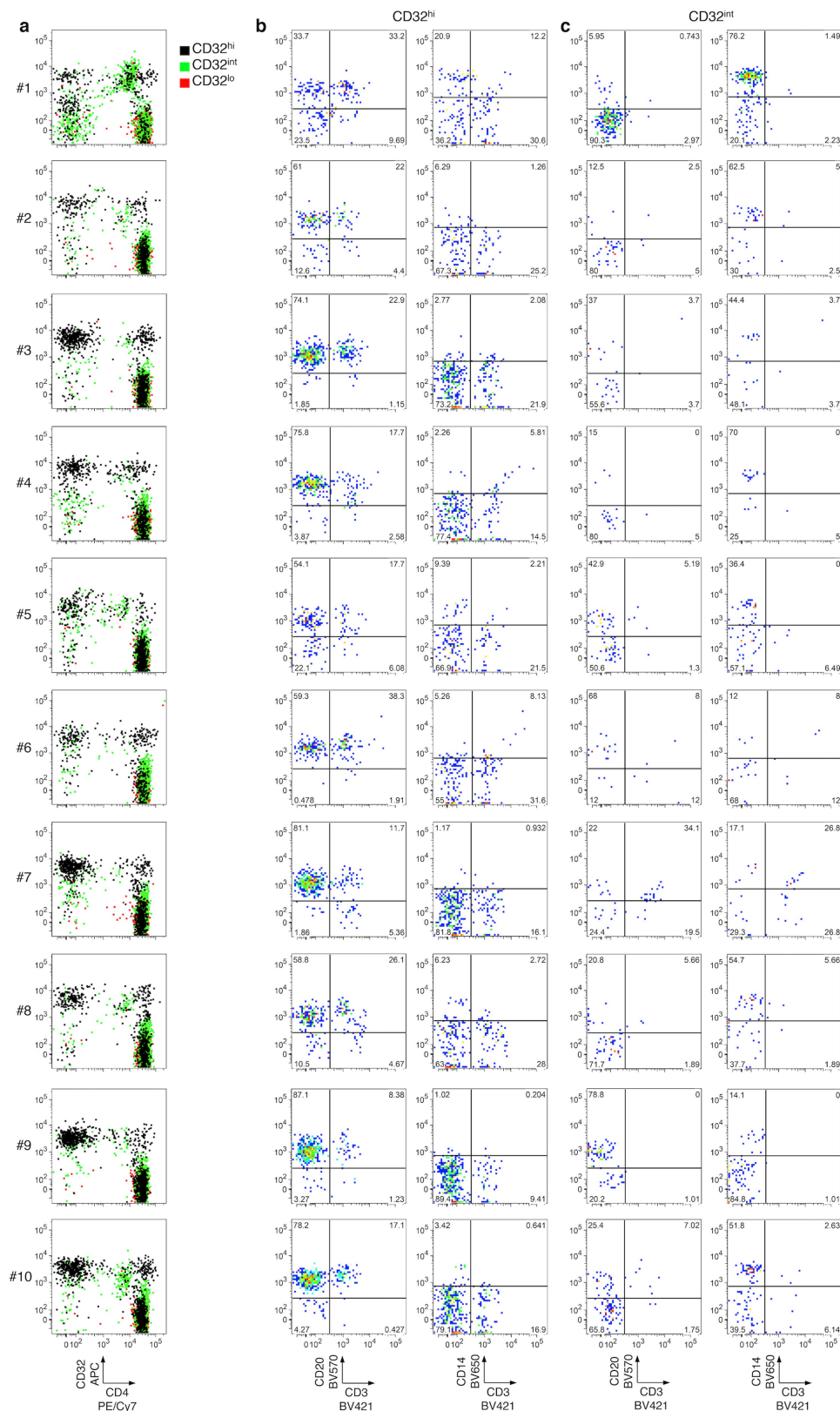
<https://doi.org/10.1038/s41586-018-0493-4>

BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 1 | Flow cytometry of CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations from PBMCs. Single lymphocytes (first two columns) that were viable (third column), CD3⁺ (fourth column), CD4⁺

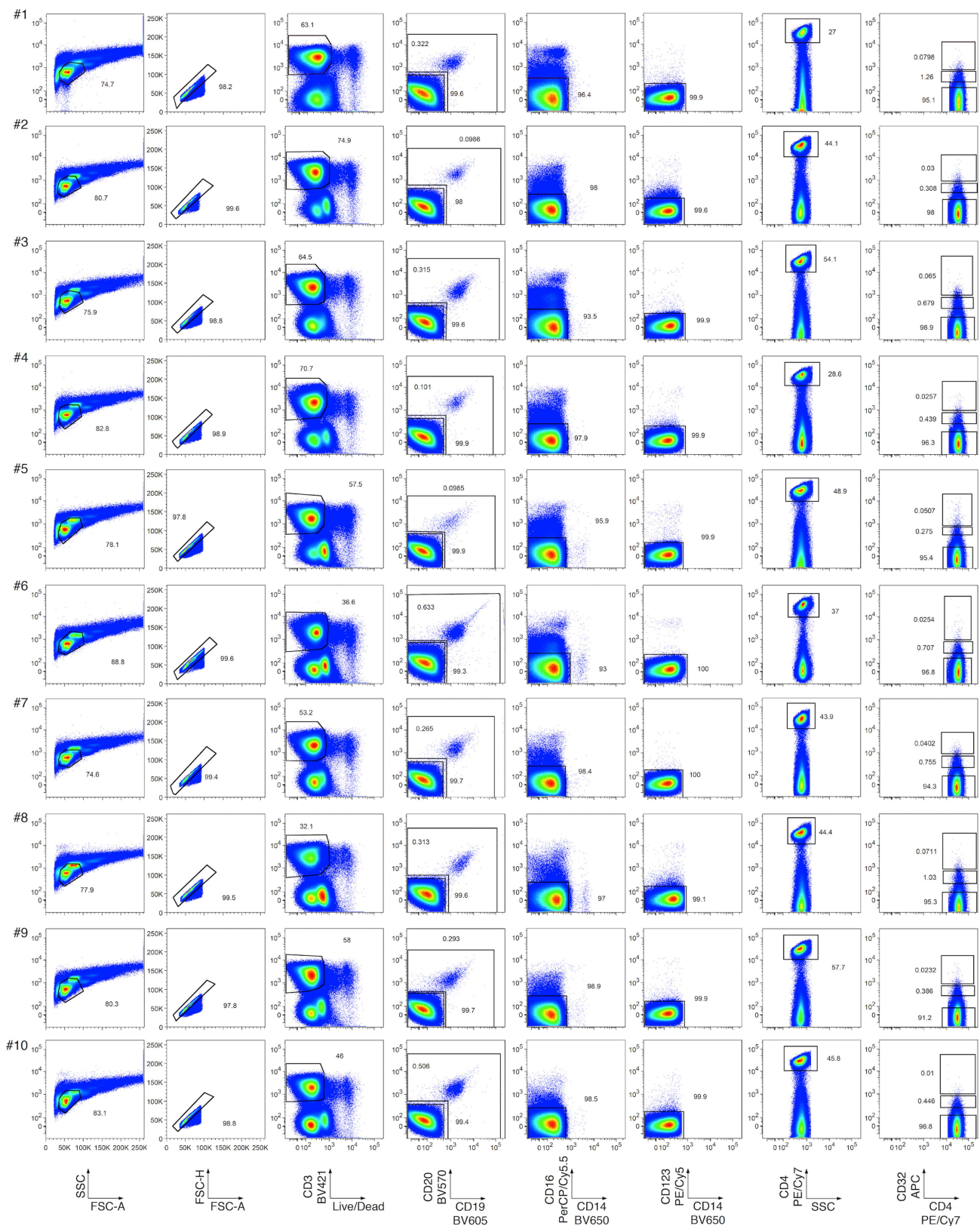
(fifth column), and CD32^{hi}, CD32^{int} or CD32^{lo} (sixth column) were sorted as described in Descours et al.¹.



Extended Data Fig. 2 | Post-sort flow cytometry of CD32⁺CD4⁺ subsets that were CD32^{hi}, CD32^{int} or CD32^{lo}. Cells were sorted as in Extended Data Fig. 1. **a**, Overlay plots of CD32 and CD4 expression by cells in CD32^{hi}, CD32^{int} and CD32^{lo} sorted populations. Note the heterogeneous pattern of cells from the CD32^{hi} and CD32^{int} populations. **b**, **c**, CD20,

CD14 and CD3 staining in the CD32⁺ cells from the CD32^{hi} (**b**) and the CD32^{int} (**c**) subsets. Note the large proportions of all CD32⁺ cells bearing surface markers consistent with B cells (CD20⁺CD3⁻) or monocytes (CD14⁺CD3⁻) after sorting.

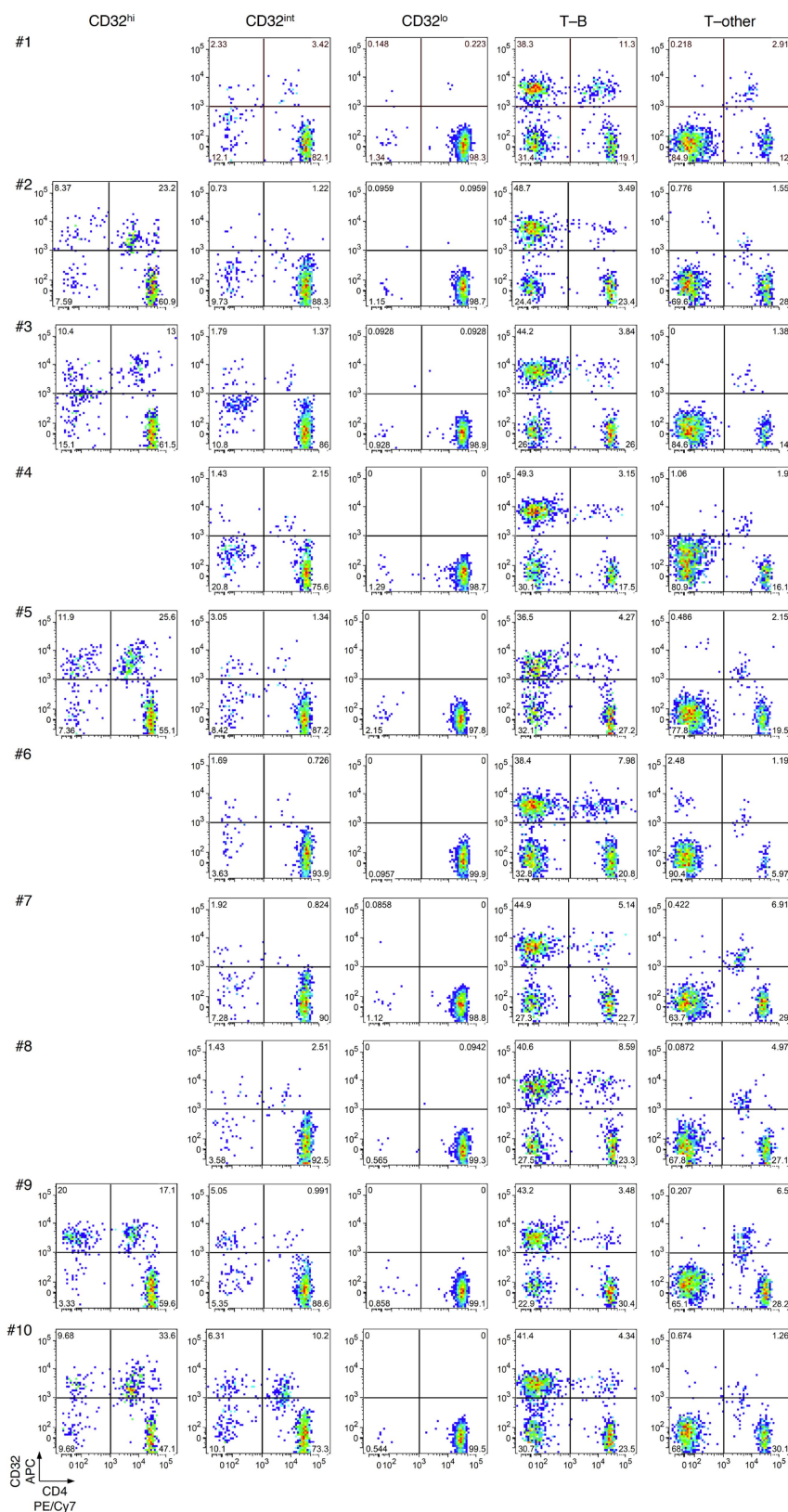
BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 3 | Flow cytometry of PBMCs sorted by alternative gating for CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations, as well as T cell populations bearing markers of B cells (T-B) or other non-CD4-T-cells (T-other). Cells in an inclusive light scatter gate consistent with either small lymphocytes or larger cells (first column) were enriched for single cells (second column). Within these gates, viable CD3⁺ cells

(third column) that were CD19⁻ and CD20⁻ (lower gate, fourth column), CD16⁻ and CD14⁻ (fifth column), CD123⁻ (sixth column), CD4⁺ (seventh column), and CD32^{hi}, CD32^{int} or CD32^{lo} were then collected. Cells that were CD3⁺ and bearing markers of B cells (T-B; upper gate, fourth column) or other non-CD4-T-cells (T-other; combined ungated events from fifth and sixth columns) were also collected in separate tubes.

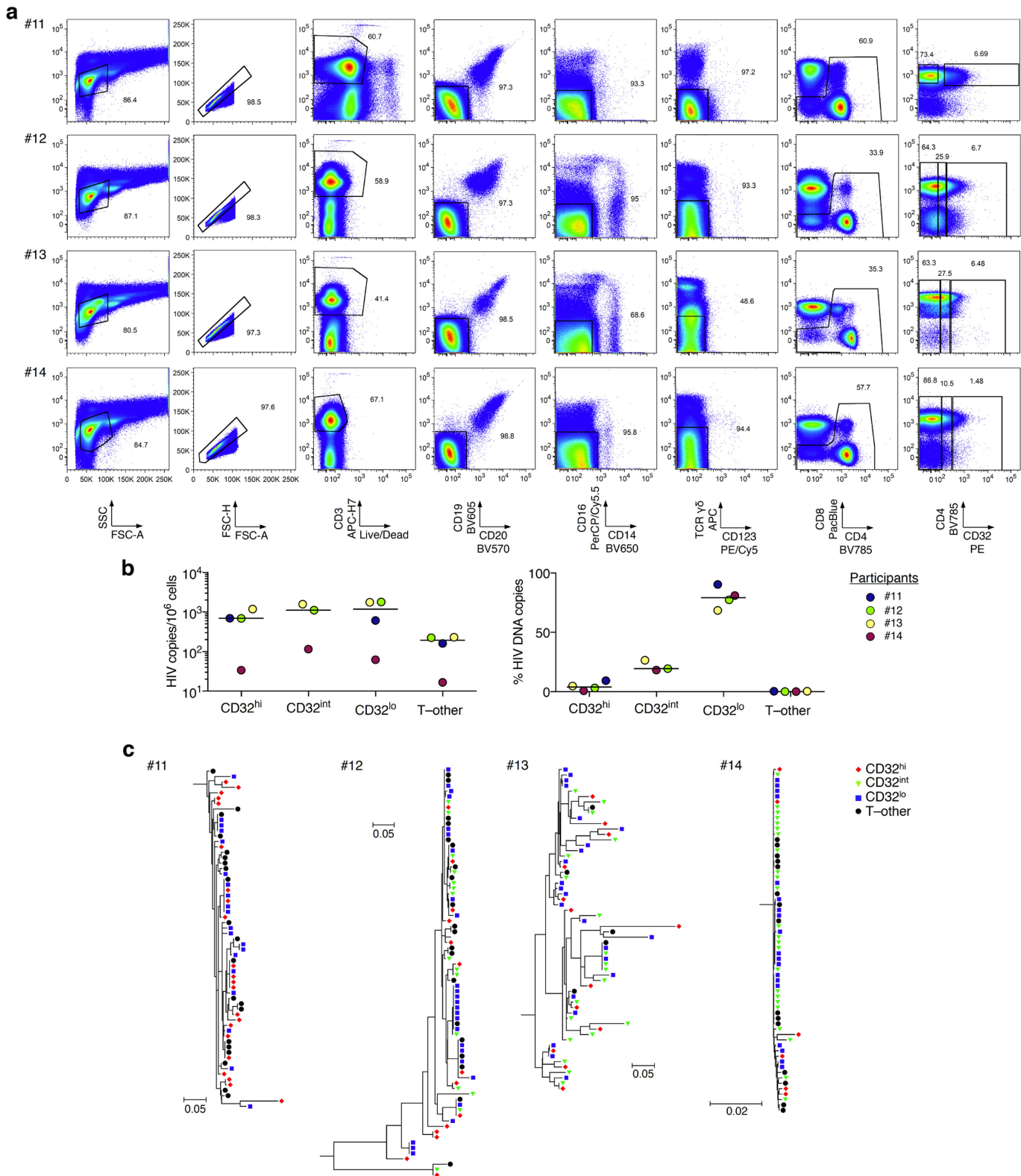
BRIEF COMMUNICATIONS ARISING



Extended Data Fig. 4 | Post-sort flow cytometry of CD32 and CD4 expression by CD32^{hi}, CD32^{int}, CD32^{lo}, T-B and T-other cell subsets. Cells were sorted as in Extended Data Fig. 3. Note the large proportions of all CD32⁺ cells that did not show high CD4 expression after sorting.

Post-sort analyses of CD3⁺CD4⁺CD32^{hi} populations were deferred in cases in which these populations were too small to permit both post-sort analysis and downstream HIV DNA quantification (that is, donors # 1, 4 and 6–8).

BRIEF COMMUNICATIONS ARISING



BRIEF COMMUNICATIONS ARISING

Extended Data Fig. 5 | Flow cytometry, HIV DNA levels, and single-copy HIV DNA sequence analysis from CD32^{hi}, CD32^{int} and CD32^{lo} CD4 T cell populations, and from T cells also bearing non-CD4-T-cell markers. **a**, PBMCs from four additional study participants were collected from whole blood by venipuncture with immediate processing (without cryopreservation). The T-other population was collected as a combination of the ungated events from CD19/CD20, CD16/CD14 and $\gamma\delta$ T cell receptor/CD123 exclusion plots (fourth, fifth and sixth columns). **b**, Left, copies of HIV DNA per million cells sorted from four additional study participants as in **a**. Right, percentages of all HIV DNA copies detected in

blood cells deriving from CD32^{hi}, CD32^{int}, CD32^{lo} and T-other subsets, calculated by adjusting values in the left panel for the relative proportions of these subsets determined using FACS data. **c**, Sequences of individual HIV DNA copies were determined by Sanger sequencing of products obtained by fluorescence-assisted clonal amplification, which amplifies a region of the HIV *env* gene. Phylogenetic trees were constructed as described in the Supplementary Methods. All Bonferroni-corrected Slatkin-Maddison *P* values for genetic compartmentalization between any two subsets were greater than 0.05 in all four participants.

The role of CD32 during HIV-1 infection

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

The persistence of latent HIV-1 in resting memory CD4⁺ T cells is a major barrier to a cure, and a biomarker for latently infected cells would be of great scientific and clinical importance^{1–5}. Using an elegant discovery-based approach, Descours et al.⁶ reported that CD32a, an Fcγ receptor not normally expressed on T cells, is a potential biomarker for the HIV-1 reservoir in CD4⁺ T cells⁶. Using a quantitative viral outgrowth assay (qVOA), we show that CD32⁺CD4⁺ T cells do not contain the majority of intact proviruses in the latent reservoir and that the enrichment found by Descours et al.⁶ may in part reflect the use of an ultrasensitive ELISA that does not predict exponential viral outgrowth. Our studies show that CD32 is not a biomarker for the major population of latently infected CD4⁺ T cells. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0496-1> (2018).

If CD32a is a biomarker for latent HIV-1 infection in CD4⁺ T cells, one that is never expressed on CD4⁺ T cells in the absence of HIV-1 infection, then a difference in the frequency of CD4⁺ T cells that express CD32 in HIV-1-infected individuals relative to the frequency in healthy donors is expected. We isolated CD4⁺ T cells from infected and uninfected donors by negative selection and analysed the expression of CD32 and CD4 by flow cytometry. In healthy donors, an average of 0.019% of CD4⁺ T cells was also CD32⁺ (Fig. 1a). This value is not significantly different from levels in HIV-1-infected individuals (Fig. 1a; average 0.011%, $P = 0.1143$) or from values previously reported by Descours et al.⁶ in HIV-1-infected individuals (0.016%, $P = 0.66$). Thus, CD32 does not seem to be a specific biomarker of latently infected CD4⁺ T cells.

To examine whether replication-competent proviruses were present in CD4⁺CD32^{hi} T cells, total CD4⁺ T cells were isolated by negative selection from six HIV-1⁺ individuals that were treated with suppressive anti-retroviral therapy (ART) for at least 6 months (Supplementary Table 1). Freshly isolated cells were stained and sorted to obtain CD4⁺CD32^{hi} and CD4⁺CD32[−] populations, which were analysed in qVOAs⁷ (Fig. 1b, protocol 1). The number of CD4⁺CD32^{hi} cells assayed for each subject is shown in Fig. 1c. On day 14, outgrowth was measured using a standard ELISA for the HIV-1 p24 antigen. CD4⁺CD32^{hi} wells from all subjects were negative for p24 on day 14, and remained negative after an additional week of culture. Conversely, outgrowth was observed in CD4⁺CD32[−] wells from all subjects on both days 14 and 21. The mean infected cell frequency, 1.37 infectious units per million cells (IUPM), was comparable to values previously measured in resting CD4⁺ T cells in several studies (0.03–3.00 IUPM in HIV-1-infected patients⁸, 0.97 IUPM in chronically infected patients⁹) and to values previously measured in the same subjects (mean value 1.33 IUPM) (Fig. 1d, Supplementary Table 2). If the enrichment of proviruses in CD32⁺ cells reported by Descours et al.⁶ was characteristic of replication-competent proviruses, then outgrowth from CD4⁺CD32^{hi} T cells should have been seen (Fig. 1e).

One possible explanation for the discrepancy between our results and those of Descours et al.⁶ is that some latent HIV-1 may be present in a previously undescribed population of CD4⁺ T cells that express CD32 together with other non-T-cell lineage markers. Such cells would be removed during the negative selection used to isolate CD4⁺ T cells. Therefore, we freshly isolated total CD4⁺ cells from infected donors on suppressive ART using two methods: negative selection to remove other lineages, leaving untouched CD4⁺ T cells, and positive selection for

cells expressing CD4 (Fig. 1b, protocol 2). Both CD4⁺ populations were analysed by qVOA. No significant differences were observed in the frequencies of latently infected cells (Fig. 1f). Furthermore, no significant differences in proviral DNA were observed between the purified cell populations (Fig. 1g). Because CD4 is required for HIV-1 entry into the host cell, cell populations obtained via positive selection for CD4 should include every latently infected CD4⁺ T cell. Given that neither the infected cell frequencies nor the levels of proviral DNA differed between the purified cell populations, we conclude that no additional sizable population of latently infected cells was recovered by positive CD4 selection.

In further studies, we used a cell sorting strategy identical to that of Descours et al.⁶ on samples freshly isolated from six subjects receiving ART treatment. Peripheral blood mononuclear cells (PBMCs) isolated from subjects were stained and sorted to obtain CD3⁺CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32[−] cell populations that were tested for latently infected cells by qVOA analysis. The numbers of CD3⁺CD4⁺CD32^{hi} cells assayed for each subject are shown in Fig. 1c and Supplementary Table 3. In addition, total CD4⁺ cells were obtained by staining PBMCs for CD4 and sorting for CD4⁺ cells (Fig. 1b, protocol 3). qVOA results showed that both the CD3⁺CD4⁺CD32[−] and the total CD4⁺ T cell populations had the same infected cell frequencies that were comparable to frequencies measured in other studies¹⁰. However, we observed no outgrowth in CD3⁺CD4⁺CD32^{hi} cultures (Fig. 1h, Supplementary Table 2).

We also analysed CD3⁺CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32[−] cells isolated by the method of Descours et al.⁶ for the presence of proviral DNA by qPCR. We found 89 copies of *gag* per million CD3⁺CD4⁺CD32[−] cells, which is similar to previous measurements in total CD4⁺ T cells¹¹. However, no proviral DNA was detected after DNA extraction from 39,000 CD3⁺CD4⁺CD32^{hi} cells and subsequent qPCR analysis (data not shown). This finding makes it highly unlikely that this cell population is enriched for HIV-1 to a level of more than one provirus copy per cell, as reported by Descours et al.⁶. We caution that the normalization of very low-level HIV-1 DNA measurements from qPCR reactions done with a low number of input cells could artificially produce apparent enrichments in HIV-1 DNA.

In a further attempt to explain the discordant qVOA results obtained in our studies and those of Descours et al.⁶, we tested whether the use of the ultra-sensitive p24 digital ELISA¹² and the low cell input can affect IUPM calculations, leading to erroneous overestimation of latent infection. qVOA culture supernatants were assayed for HIV-1 p24 using the ultrasensitive SIMOA p24 2.0 assay (Quanterix) on days 5, 9, 14 and 21. Using the lower limit of quantification (0.01 pg ml^{−1}) as the cut-off level, we found that two out of three qVOAs containing CD4⁺CD32^{hi} cells tested positive for p24 by this assay, even though the same wells were negative by standard ELISA, which is several orders of magnitude less sensitive (Fig. 2a). Exponential outgrowth is the hallmark of replication-competent viruses. In qVOA cultures of CD4⁺CD32[−] cells, only a fraction of the wells that were positive by SIMOA showed exponential outgrowth as determined by standard ELISA on day 21 (Fig. 2b). Importantly, CD4⁺CD32^{hi} culture wells that tested positive by SIMOA p24 assay showed no exponential outgrowth and had significantly lower levels of p24 (Fig. 2c). It is possible that low positive SIMOA values could reflect an assay artefact or the presence of defective proviruses that are still capable of producing low levels of Gag¹³. A further concern

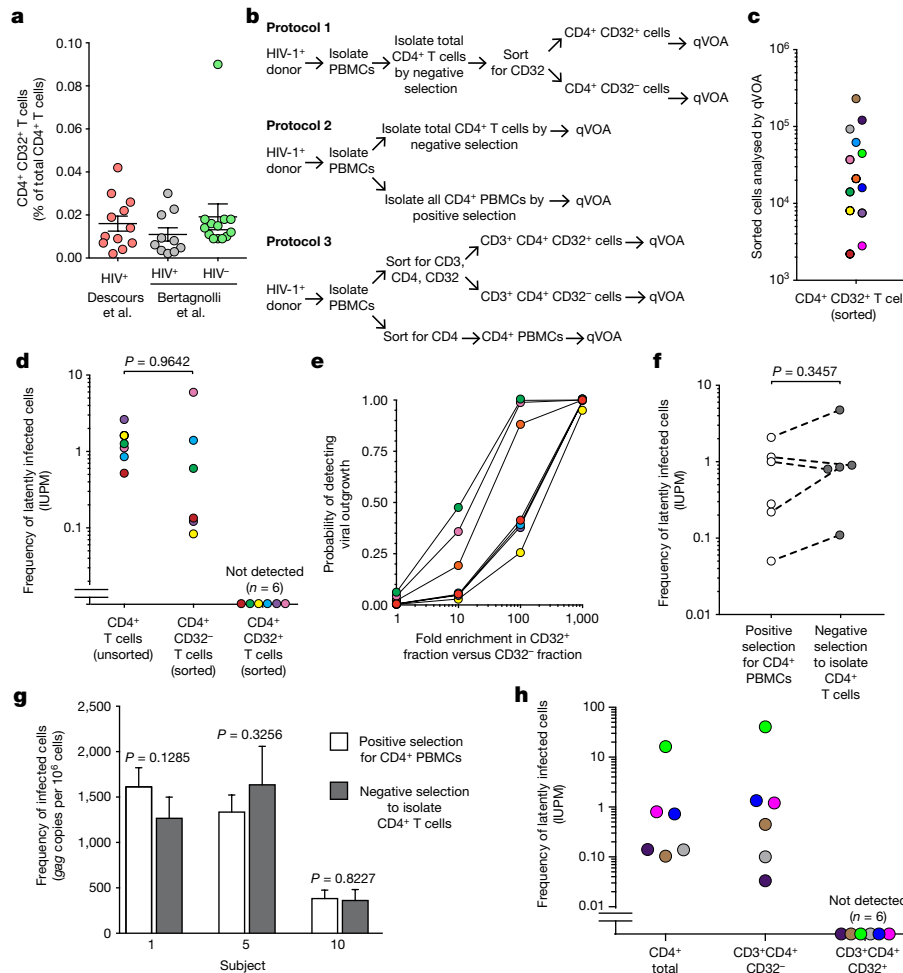


Fig. 1 | Analysis of CD4⁺CD32⁻ and CD4⁺CD32⁺ populations by qVOA and proviral DNA measurements. **a**, Percentage of CD4⁺CD32^{hi} T cells relative to total CD4⁺ T cells in healthy donors and HIV-1-infected donors. Infected donor values were obtained from supplementary table 4 of Descours et al.⁶. LLOQ, lower limit of quantification. **b**, Schematic depicting the three strategies (protocols 1–3) used to obtain different populations of CD4⁺ T cells analysed in qVOAs. **c**, Numbers of sorted CD4⁺CD32^{hi} and CD3⁺CD4⁺CD32^{hi} T cells from each subject analysed in qVOAs. **d**, Frequencies of latently infected cells among CD4⁺CD32⁻ T cells and CD4⁺CD32⁺ T cells and among total CD4⁺ T cells from the same subjects previously measured in separate experiments. Cells were isolated using protocol 1 (colours correspond to subject values from panel c). **e**, Probability of detecting outgrowth based on measured frequencies of latently infected cells among the CD4⁺CD32⁻ fraction and number of CD4⁺CD32^{hi} cells plated assuming various degrees of enrichment of HIV-1 in CD32^{hi} cells. **f**, Frequencies of latently infected cells measured in qVOAs using positive or negative selection to obtain total CD4⁺ cells (protocol 2; positive selection was accomplished by either sorting or CD4 microbead strategies, with similar results). **g**, Comparison of proviral DNA measurements obtained with qPCR on total CD4⁺ cells purified using positive or negative selection (protocol 2). **h**, Frequencies of latently infected cells among total CD4 cells, and CD3⁺CD4⁺CD32⁻ and CD3⁺CD4⁺CD32^{hi} populations. Cells were isolated using protocol 3 (colours correspond to subject values from panel c).

e, Probability of detecting outgrowth based on measured frequencies of latently infected cells among the CD4⁺CD32⁻ fraction and number of CD4⁺CD32^{hi} cells plated assuming various degrees of enrichment of HIV-1 in CD32^{hi} cells. **f**, Frequencies of latently infected cells measured in qVOAs using positive or negative selection to obtain total CD4⁺ cells (protocol 2; positive selection was accomplished by either sorting or CD4 microbead strategies, with similar results). **g**, Comparison of proviral DNA measurements obtained with qPCR on total CD4⁺ cells purified using positive or negative selection (protocol 2). **h**, Frequencies of latently infected cells among total CD4 cells, and CD3⁺CD4⁺CD32⁻ and CD3⁺CD4⁺CD32^{hi} populations. Cells were isolated using protocol 3 (colours correspond to subject values from panel c).

is that the IUPM calculations are based on cell input, fold dilutions and technical replicates¹⁴, and thus, qVOA analyses performed with very small numbers of sorted CD4⁺CD32^{hi} cells can markedly skew the frequency of cells harbouring replication-competent proviruses (five-fold dilutions from 800 to 1 cell in Descours et al.⁶). When we applied the results obtained with the SIMOA p24 assay, IUPM values ranged from 0 to 3,134 and 554 (patients 4 and 5, respectively; Fig. 2d). As a consequence, when we calculated the 'fold enrichment' of IUPM in the CD4⁺CD32^{hi} cells compared to the CD4⁺CD32⁻ cells, we observed a mean fold enrichment of 665 (range 152–1179, from the two patients with positive p24 using SIMOA), similar to what was reported by Descours et al.⁶ (Fig. 2e).

In summary, we find no evidence that CD32 expression indicates the presence of latent HIV-1, and demonstrate that at least a substantial fraction of the HIV-1 latent reservoir is in CD3⁺CD4⁺CD32⁻

T cells. Although no outgrowth could be found in cultures containing CD4⁺CD32^{hi} T cells, viral outgrowth comparable to historical measurements was found in cultures containing CD4⁺CD32⁻ T cells. The use of an ultrasensitive p24 ELISA assay may account for the apparent enrichment observed in culture experiments by Descours et al.⁶. In short, our results have demonstrated that CD32 does not define the HIV-1 reservoir and that future research is needed to identify biomarkers for latently infected cells.

We thank the study participants without whom this research would not be possible. Funding was provided by the US National Institutes of Health (NIH) Martin Delaney I4C, Beat-HIV and DARE Collaboratories by the Johns Hopkins Center for AIDS Research (P30AI094189), by NIH grant 43222, and by the Howard Hughes Medical Institute and the Bill and Melinda Gates Foundation.

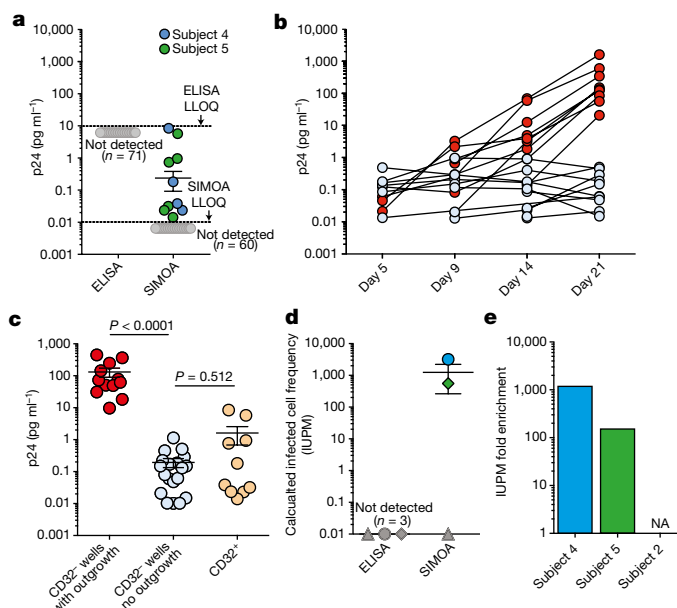


Fig. 2 | Ultrasensitive p24 measurements. **a**, Levels of p24 from CD32⁺ culture wells measured by ELISA and SIMOA (lower limit of quantification: 5–10 pg ml⁻¹ and 0.01 pg ml⁻¹, respectively) (data collected from three subjects, for a total of 71 wells). **b**, Longitudinal levels of p24 measured by SIMOA in individual culture wells in the qVOA for CD32⁻ cells from subject 5, showing wells with and without viral outgrowth (red and blue circles, respectively). **c**, Levels of p24 measured by ELISA in CD32⁻ wells with outgrowth compared with SIMOA measurements in wells with no outgrowth and CD32⁺ wells (data collected from subjects 2, 4 and 5). *P* values were determined with a non-parametric *t*-test. **d**, IUPM calculation based on ELISA and SIMOA analysis. Symbols in dark grey represent values below the limit of detection. **e**, Fold enrichment of IUPM in CD32⁺ cells (from subjects 2, 4 and 5). NA, not applicable.

Methods

qVOAs isolated CD4⁺ T cells using negative depletion and were sorted for CD32⁺ cells (Fig. 1b, protocol 1). To test whether negative depletion was causing a loss of CD32⁺ CD4⁺ T cells, outgrowth and proviral DNA were compared from qVOAs in which CD4⁺ T cells were isolated using positive selection to measurements using negative depletion. Outgrowth measurements and proviral DNA were also measured using the methods described by Descours et al.⁶. Proviral DNA measurements were performed using qPCR¹⁵. HIV-1 p24 values were measured using both a standard ELISA for p24 antigen (Perkin Elmer) and SIMOA (Quanterix). Further details are provided in Supplementary Methods.

Data availability. All data are available from the corresponding author upon reasonable request.

Lynn N. Bertagnoli¹, Jennifer A. White¹, Francesco R. Simonetti¹, Subul A. Beg^{1,2}, Jun Lai^{1,2}, Costin Tomescu³, Alexandra J. Murray¹, Annukka A. R. Antar¹, Hao Zhang⁴, Joseph B. Margolick⁴, Rebecca Hoh⁵, Stephen G. Deeks⁵, Pablo Tebas⁶, Luis J. Montaner³, Robert F. Siliciano^{1,2*}, Gregory M. Laird¹ & Janet D. Siliciano¹

¹Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ²Howard Hughes Medical Institute, Baltimore, MD, USA. ³The Wistar Institute, Philadelphia, PA, USA. ⁴Department of Molecular Microbiology and Immunology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. ⁵Division of HIV, Infectious Diseases and Global Medicine, University of California, San Francisco, CA, USA. ⁶Division of Infectious Diseases, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. *e-mail: rsiliciano@jhmi.edu

Received: 29 September 2017; Accepted: 3 April 2018;

Published online 19 September 2018.

1. Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
2. Chun, T. W. et al. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
3. Wong, J. K. et al. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
4. Richman, D. D. et al. The challenge of finding a cure for HIV infection. *Science* **323**, 1304–1307 (2009).
5. Deeks, S. G. et al. Towards an HIV cure: a global scientific strategy. *Nat. Rev. Immunol.* **12**, 607–614 (2012).
6. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
7. Laird, G. M., Rosenbloom, D. I., Lai, J., Siliciano, R. F. & Siliciano, J. D. Measuring the frequency of latent HIV-1 in resting CD4⁺ T cells using a limiting dilution coculture assay. *Methods Mol. Biol.* **1354**, 239–253 (2016).
8. Siliciano, J. D. et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4⁺ T cells. *Nat. Med.* **9**, 727–728 (2003).
9. Eriksson, S. et al. Comparative analysis of measures of viral reservoirs in HIV-1 eradication studies. *PLoS Pathog.* **9**, e1003174 (2013).
10. Crooks, A. M. et al. Precise quantitation of the latent HIV-1 reservoir: implications for eradication strategies. *J. Infect. Dis.* **212**, 1361–1365 (2015).
11. Besson, G. J. et al. HIV-1 DNA decay dynamics in blood during more than a decade of suppressive antiretroviral therapy. *Clin. Infect. Dis.* **59**, 1312–1321 (2014).
12. Passaes, C. P. & Sáez-Cirión, A. HIV cure research: advances and prospects. *Virology* **454–455**, 340–352 (2014).
13. Pollack, R. A. et al. Defective HIV-1 proviruses are expressed and can be recognized by cytotoxic T lymphocytes, which shape the proviral landscape. *Cell Host Microbe* **21**, 494–506.e4 (2017).
14. Rosenbloom, D. I. et al. Designing and interpreting limiting dilution assays: general principles and applications to the latent reservoir for human immunodeficiency virus-1. *Open Forum Infect. Dis.* **2**, ofv123 (2015).
15. Massanella, M., Gianella, S., Lada, S. M., Richman, D. D. & Strain, M. C. Quantification of total and 2-LTR (long terminal repeat) HIV DNA, HIV RNA and herpesvirus DNA in PBMCs. *Bio Protoc.* **5**, e1492 (2015).

Author contributions L.N.B., J.A.W., G.M.L., F.R.S. and J.D.S. designed experiments. S.A.B., C.T. and L.J.M. obtained samples. L.N.B., J.A.W., S.A.B., G.M.L., F.R.S., J.L., A.J.M., A.A.R.A. and J.D.S. performed experiments. F.R.S., H.J. and J.B.M. performed cell sorting. L.N.B., J.A.W., F.R.S., A.J.M., A.A.R.A., R.F.S. and J.D.S. analysed the data and wrote the manuscript.

Competing interests Declared none.

Additional information

Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.F.S.

<https://doi.org/10.1038/s41586-018-0494-3>

Evidence that CD32a does not mark the HIV-1 latent reservoir

ARISING FROM B. Descours et al. *Nature* **543**, 564–567 (2017); <https://doi.org/10.1038/nature21710>

A recent report by Descours et al.¹ suggests that the cell surface expression of the low affinity Fc receptor CD32a (also known as FcγRIIa) marks the replication-competent HIV-1 reservoir in CD4⁺ T cells from 12 HIV-1-infected participants receiving suppressive anti-retroviral therapy (ART)¹. We have undertaken considerable efforts to replicate these findings using peripheral blood mononuclear cells (PBMCs) from 20 HIV-1-infected, ART-suppressed participants (Extended Data Table 1). We found no evidence to suggest that CD32a marks a CD4⁺ T cell population enriched in either HIV-1 DNA or replication-competent HIV-1 in our study participants. There is a Reply to this Comment by Descours, B. et al. *Nature* **561**, <https://doi.org/10.1038/s41586-s41586-018-0496-1> (2018).

To validate these findings, we adopted the same gating strategy as described by Descours et al.¹ to define CD4⁺ T cell populations (Supplementary Fig. 1a). The CD32 antigen was identified using the same antibody clone (FUN-2) as described by Descours et al.¹. We observed the same CD4⁺ T cell subsets that stained at a high cell surface density of CD32 (CD4⁺CD32^{high}), an intermediate cell surface density of CD32 (CD4⁺CD32^{int}), and a CD4⁺ T cell subset lacking CD32 expression (CD4⁺CD32^{neg}). We obtained frequencies of CD4⁺CD32^{high} T cells that ranged from 0.002% to 0.026%, with a median value (0.012%) that was identical to that reported by Descours et al.¹ (Extended Data Table 2 and Supplementary Fig. 1a). Notably, we confirmed that this same CD4⁺CD32^{high} population is also present in PBMCs isolated from eight healthy donors and exists at similar frequencies to that in HIV-1-infected samples ($P = 0.971$, Extended Data Fig. 1a).

Next, we assessed the amount of replication-competent HIV-1 isolated from the same 20 participants by measuring the infectious unit per million cells (IUPM) in CD4⁺ T cells (range 0.01–37.5, median 0.46). Participant CD4⁺CD32^{high} T cell populations were colour-coded in descending order, and then divided into quartiles that corresponded to the relative frequency of CD4⁺CD32^{high} cells present in these samples (Fig. 1a).

After cytometric sorting of the various CD4⁺CD32 subsets, we quantified HIV-1 DNA in each population (total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high}, Fig. 1b) using droplet digital PCR (ddPCR), as described in the Methods. We found no evidence of HIV-1 DNA enrichment in the CD4⁺CD32^{high} fraction. We observed no significant difference in HIV-1 DNA between any populations and the CD4⁺CD32^{high} T cell population ($P = 0.28$). In fact, levels of HIV-1 DNA in the CD4⁺CD32^{high} T cell subsets isolated from nine participants was at the assay limit of detection (Fig. 1b). After correction for cell input in the CD4⁺CD32^{high} fraction, as estimated DNA values, we saw no evidence for HIV-1 DNA enrichment (open symbols in Extended Data Fig. 1b).

We then compared the relative frequency of the CD4⁺CD32^{high} T cell populations and the viral replicative capacity (IUPM values) per participant, but no relationship between the two parameters was observed (Fig. 1c). All values have been tabulated in Extended Data Table 2.

The HIV-1 reservoir largely resides in quiescent CD4⁺ T cells^{2,3}. Therefore, we sought to confirm the activation status of the CD4⁺ T cell populations by measuring the frequency of the activation markers CD69, CD25 and HLA-DR on CD4⁺ T cell subsets from all

participants. We found that the CD4⁺CD32^{high} T cells were highly activated compared to the CD4⁺CD32^{neg} T cells ($P < 0.0001$). Notably, among the activation markers, HLA-DR was particularly enriched,

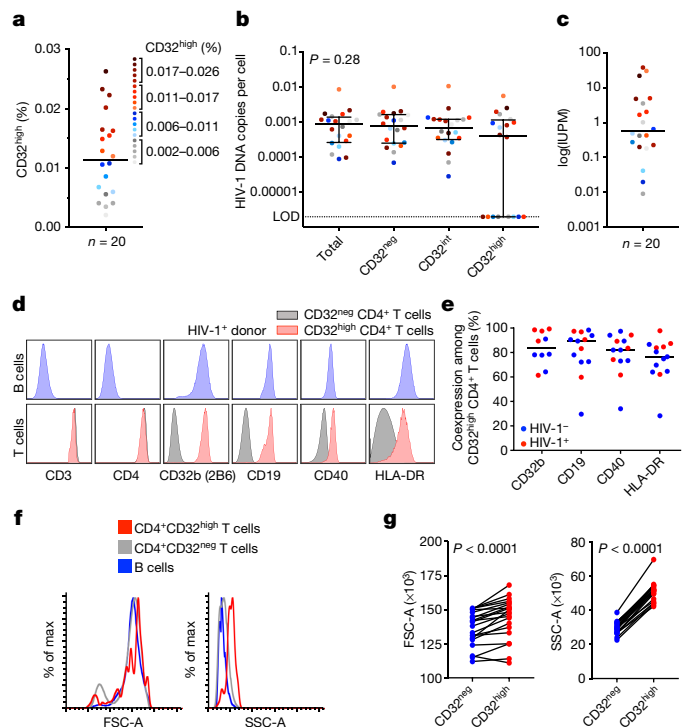


Fig. 1 | CD32-expressing CD4⁺ T cells are not enriched in HIV-1 DNA and express markers of B cell origin. **a–c**, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells from PBMCs of ART-suppressed, HIV-1-infected patients ($n = 20$) were sorted, and HIV-1 DNA was measured by ddPCR. **a**, Dividing the frequency (in percentage) of CD4⁺CD32^{high} T cells from all participants into quartiles, the values are shown as below or above the median. **b**, DNA copies per cell in sorted subsets of total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells are shown, with median and interquartile range (IQR). P value determined by Kruskal–Wallis test. LOD, limit of detection. **c**, IUPM in CD4⁺ T cells of each participant is shown in the colour corresponding to its frequency of CD4⁺CD32^{high} cells in panel **a**. **d**, **e**, CD32^{neg} and CD32^{high} (identified using FUN-2) CD4⁺ T cells from human PBMCs were assessed by flow cytometry for the expression of CD32b (2B6 antibody), CD19, CD40 and HLA-DR and compared to B cells (CD3⁺CD14⁺CD19⁺ lymphocytes). **d**, Representative flow cytometry results per cell antigen levels on B cells (top, blue histograms) and on CD32^{neg} and CD32^{high} CD4⁺ T cells (bottom, grey and red histograms, respectively) from PBMCs from an HIV-1⁺ participant. **e**, Frequency of CD4⁺CD32^{high} T cells staining positive for CD32b (2B6), CD19, CD40 or HLA-DR from HIV-1⁺ ($n = 5$) and HIV-1[−] ($n = 5–8$) human donor PBMC samples. Bars denote median values. **f**, Representative histograms of the FSC-A and SSC-A of B cells and CD32^{neg} and CD32^{high} CD4⁺ T cells from PBMCs of an HIV-1⁺, ART-suppressed participant sorted on a BD FACSAria II. **g**, Comparisons of the median FSC-A and SSC-A values between CD32^{neg} and CD32^{high} CD4⁺ T cell subsets from HIV-1⁺, ART-suppressed participants ($n = 20$). P values were determined using a paired t -test.

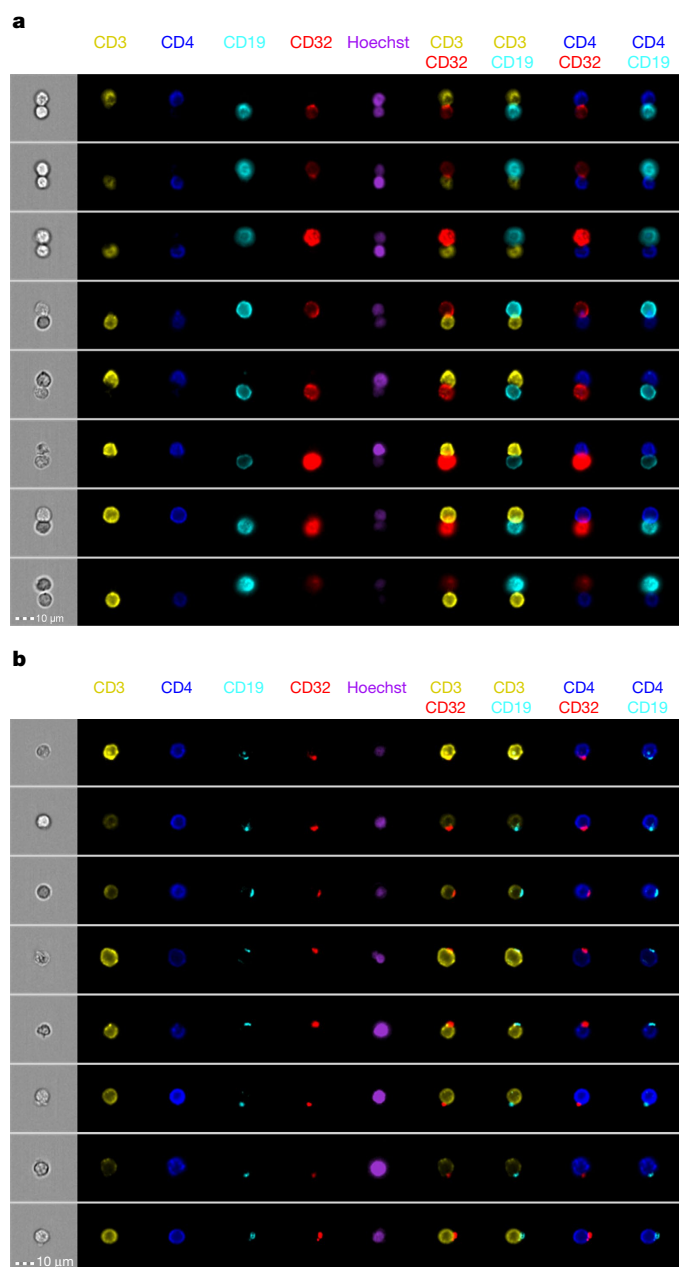


Fig. 2 | Flow cytometry imaging of sorted CD32-expressing T cells. **a**, Representative bright-field and pseudo-colour fluorescence images of T–B cell conjugates found in the CD32^{high} CD4⁺ T cell population sorted from PBMCs from HIV-1⁺, ART-suppressed participants, and imaged using Amnis technology. **b**, Representative images of punctate CD32 staining found on single T cells in the CD32^{high} and CD32^{int} population sorted from HIV-1⁺, ART-suppressed participant PBMCs.

marking approximately 75% of all CD4⁺CD32^{high} cells (median 74%), compared to CD4⁺CD32^{neg} cell populations (median 1.4%) (Extended Data Fig. 1c, $P < 0.0001$).

Two CD32 isoforms (CD32a and CD32b) are known to be expressed among all antigen presenting cells (APCs), but not typically on T cells. Therefore, we sought to exclude any APCs as potential contaminants of flow cytometry sorting. We evaluated the co-expression of lineage markers for all major CD32-bearing cells including monocytes, B cells, dendritic cells, granulocytes and natural killer cells. As expected, all CD32⁺ T cells expressed high amounts of CD3 and CD4 (Fig. 1d). However, we found that most CD4⁺CD32^{high} T cells from HIV-1⁺

patients, and also from healthy donors, co-expressed several B cell markers including CD19, CD40 and HLA-DR (Fig. 1d, e, Extended Data Fig. 2a). Notably, the B cell antigens found on CD4⁺CD32^{high} T cells were present at similar cell-surface densities as detected on bona fide B cells (Fig. 1d, e, Extended Data Fig. 2a).

The demonstration that the CD4⁺CD32^{high} fraction seen in HIV-1⁺ patients was marked with several B cell antigens and was similarly present in naive donors led us to investigate the origin of these B cell markers on a CD4⁺ T cell. Several reports have shown that B cells exclusively express the CD32b isoform⁴. The FUN-2 antibody clone used by Descours et al.¹ cannot distinguish between the CD32a and CD32b isoforms. Therefore, we used the monoclonal antibody clone 2B6 that has been reported to exclusively bind to CD32b^{4,5}. After co-staining PBMCs from HIV-1⁺ and HIV-1⁻ individuals with both the FUN-2 and the 2B6 antibodies, we found that all CD4⁺CD32^{high} T cells were marked only by the CD32b isoform and not by CD32a (Fig. 1d, e, Extended Data Fig. 2b), indicating that B cells are the origin of the CD32b antigen that marks the CD4⁺CD32^{high} T cells.

We sought to confirm this by determining whether *CD32A* (also known as *FCGR2A*) or *CD32B* (*FCGR2B*) mRNA was endogenously produced in the CD4⁺CD32^{high} subsets. After isolating total cellular RNA from various sorted T cell subsets, we used established reverse transcription PCR (RT–PCR) primers and probes that are specific to the CD32a and CD32b isoforms, as described in the Methods. We found that sorted CD4⁺CD32^{high} T cells from four HIV-1-infected participants did not contain detectable levels of the *CD32A* isoform. However, the *CD32B* mRNA isoform was readily detected in CD4⁺CD32^{high} T cells isolated from two out of four HIV-1⁺ patients (Extended Data Fig. 2c). By additional RT–PCR analysis, we detected both *CD3G* and *CD19* transcripts in the same CD4⁺CD32^{high} T population, indicating that the CD32b marking the CD4⁺CD32^{high} T cells may be from B cells expressing cognate CD32b (Extended Data Fig. 2d).

Because this may require cell-to-cell interaction, we performed a back-gating analysis of our flow cytometry data and confirmed that all CD4⁺CD32^{high} populations were identified within single-cell gates (Supplementary Fig. 1b). However, post-hoc analysis comparing the forward and side scatter light pulse area (FSC–A and SSC–A, respectively) values between CD4⁺CD32^{neg} and CD4⁺CD32^{high} T cells showed that the CD4⁺CD32^{high} populations had both a significantly higher FSC–A ($P < 0.0001$) and SSC–A ($P < 0.0001$), suggesting that the CD4⁺CD32^{high} population may consist largely of cell doublets (Fig. 1f, g).

We next used Amnis imaging flow cytometry to visualize the sorted CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} cell populations directly. As expected, the CD4⁺CD32^{neg} and the CD4⁺CD32^{int} cell populations each consisted of more than 99% single cells. However, the CD4⁺CD32^{high} fraction contained a high frequency of cell doublets (mean value 94%) (Extended Data Fig. 3). Of these ‘doublets’, approximately 70% seemed to be coincident doublets, and 30% were conjugates of T and B cells (Fig. 2a and Extended Data Fig. 3b).

We observed no examples in which CD32 staining on T cells was distributed throughout the cell membrane, supporting the idea that the CD32 found in the CD4⁺CD32^{high} population is not the result of endogenous expression from CD4⁺ T cells. Of the instances in which CD32 was detected on a T cell in the CD4⁺CD32^{high} population, the staining was punctate and often co-localized with punctate CD19 staining (Fig. 2b), suggesting that CD32 was acquired via contact between B and T cells. We noted that the frequency of T cells with punctate CD32 staining was substantially higher in the sorted CD32^{int} population. Thus, sorting for CD4⁺ T cells with a ‘high’ surface density of CD32 results in the selective enrichment of contaminating T–B cell doublets. As shown in Supplementary Fig. 1, these doublets cannot be discerned by routine cytometric FSC and SSC singlet gating strategies.

In summary, using samples from 20 HIV-1-infected, ART-suppressed participants, our data contradict the assertion that CD32a is a marker of

the replication-competent viral reservoir. Although we did detect similar frequencies of CD4⁺CD32^{high} populations to Descours et al.¹, we found no difference in the total HIV-1 DNA content between CD4⁺ T cell populations including or excluding the CD32^{high} fractions (Fig. 1b).

Notably, the CD4⁺CD32^{high} population was highly activated. Previous studies that have evaluated CD32 expression on T cells suggest that it may be detected after activation^{6,7} and led us to believe that this population may be atypical compared to a quiescent population harbouring the HIV-1 reservoir^{2,3}.

Our additional findings are incongruent with CD32a marking the replication-competent reservoir in CD4⁺ T cells; our phenotyping and RT-PCR experiments indicate that it is the CD32b isoform that marks the CD4⁺CD32^{high} cells (Fig. 1d, e, Extended Data Fig. 2b–d). This finding, combined with the demonstration that this cell population is found in uninfected individuals, conflicts with the assertion of Descours et al.¹ that CD32a is upregulated after the establishment of viral latency. Recent reports have corroborated the absence of CD32a transcripts in reactivated, clonal HIV-1-infected CD4⁺ T cells⁸.

The surface density of CD32b (and other B cell markers) on the CD4⁺CD32^{high} population was observed at similar densities to that on B cells. These data, combined with the post-hoc analysis, suggests that this population may be largely comprised of doublets. Direct interrogation of the CD4⁺CD32^{high} population via Amnis imaging confirmed that this population consisted largely of contaminating doublets; either co-incident events or cell-to-cell conjugates (Fig. 2a).

We demonstrate that the mechanism by which the CD32b isoform labels the CD4⁺CD32^{high} populations is through the direct interaction of CD4⁺ T and B cells, and possible trogocytotic transfer of B cell antigens to T cells, as observed in the CD4⁺CD32^{int} population (Fig. 2b). This may explain the transfer or membrane painting of antigens such as CD32b, CD40 and HLA-DR, among other markers^{9–11}. Not only have cell-to-cell membrane transfers been shown to occur commonly *in vivo* during viral infections, but such transfers largely occur on activated cells¹². Membrane-bound Fcγ receptors, including CD32b, are known to be extracted from APCs and then transferred to T cells, and serve as a surrogate of recent T cell and APC interactions¹³. Our demonstration of T–B cell conjugates in the CD4⁺CD32^{high} population and high levels of single cells in the CD4⁺CD32^{int} population support this notion (Fig. 2a, b).

Collectively, our findings confirm that selectively sorting for T cells with a high surface density of CD32 results in the enrichment of T–B cell doublet contaminants, which cannot be discerned by routine gating strategies. The true isoform, CD32b, that marks the CD4⁺CD32^{high} population is probably indicative of dynamic CD4⁺ T cell interaction with B cells, rather than a marker of the HIV-1 reservoir^{14,15}.

We thank S. Mordecai for Amnis technical expertise, and acknowledge support from NIAID grants AI091514, AI122942, AI127089 and AI131365 awarded to J.B.W. Support was also provided by the NIAID awarded Martin Delaney Collaboratory ‘BELIEVE’ grant AI126617, co-funded by NIDA, NIMH and NINDS awarded to D.F.N.

Methods

HIV-1⁺ participants were recruited through: The Maple Leaf Medical clinic in Toronto, Canada; The HIV Eradication and Latency (HEAL) cohort of Brigham and Women's and Massachusetts General Hospital; The Whitmann Walker Clinic in Washington, DC; or the Hospital of the University of Pennsylvania. The study was approved by the University of Toronto, The University of Pennsylvania and George Washington University ethics committees and according to the protocol approved by the Partners Human Research Committee and Institutional Review Board (IRB). Written informed consent was obtained from each participant.

The percentage of CD32⁺ (clone FUN-2) CD4⁺ T cells was measured in samples from study participants. Both CD32⁺ and CD32^{neg} CD4⁺ T cells were sorted and viral DNA was measured using ddPCR. The analysis of cell lineage markers by flow cytometry and RT-PCR was also conducted. Flow cytometry sorts from PBMCs used in HIV-1 DNA analyses were performed on cell subsets and assessed using Amnis imaging flow cytometry.

Data availability. All data and reagents are available from the corresponding author upon request.

Christa E. Osuna¹, So-Yon Lim¹, Jessica L. Kublin¹, Richard Apps², Elsa Chen¹, Talia M. Mota³, Szu-Han Huang³, Yanqin Ren³, Nathaniel D. Bachtel³, Athe M. Tsibris⁴, Margaret E. Ackerman⁵, R. Brad Jones³, Douglas F. Nixon³ & James B. Whitney^{1,6*}

¹Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. ²Center for Human Immunology, National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA. ³Division of Infectious Diseases, Weill Department of Medicine, Weill Cornell Medical College, New York, NY, USA. ⁴Brigham and Women's Hospital, Boston, Massachusetts Harvard Medical School, Boston, MA, USA. ⁵Thayer School of Engineering, Dartmouth College, Hanover, NH, USA. ⁶Ragon Institute of MGH, MIT, and Harvard, Cambridge, MA, USA. *e-mail: jwhitne2@bidmc.harvard.edu

Received: 11 August 2017; Accepted: 24 May 2018;
Published online 19 September 2018.

- Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
- Chun, T. W. et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387**, 183–188 (1997).
- Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
- Veri, M. C. et al. Monoclonal antibodies capable of discriminating the human inhibitory Fcγ-receptor IIB (CD32B) from the activating Fcγ-receptor IIA (CD32A): biochemical, biological and functional characterization. *Immunology* **121**, 392–404 (2007).
- Boruchov, A. M. et al. Activating and inhibitory IgG Fc receptors on human DCs mediate opposing functions. *J. Clin. Invest.* **115**, 2914–2923 (2005).
- Engelhardt, W., Matzke, J. & Schmidt, R. E. Activation-dependent expression of low affinity IgG receptors FcγRII(CD32) and FcγRIII(CD16) in subpopulations of human T lymphocytes. *Immunobiology* **192**, 297–320 (1995).
- Sandilands, G. P. et al. Differential expression of CD32 isoforms following alloactivation of human T cells. *Immunology* **91**, 204–211 (1997).
- Cohn, L. B. et al. Clonal CD4⁺ T cells in the HIV-1 latent reservoir display a distinct gene profile upon reactivation. *Nat. Med.* **24**, 604–609 (2018).
- Cone, R. E., Sprent, J. & Marchalonis, J. J. Antigen-binding specificity of isolated cell-surface immunoglobulin from thymus cells activated to histocompatibility antigens. *Proc. Natl Acad. Sci. USA* **69**, 2556–2560 (1972).
- Hwang, I. et al. T cells can use either T cell receptor or CD28 receptors to absorb and internalize cell surface molecules derived from antigen-presenting cells. *J. Exp. Med.* **191**, 1137–1148 (2000).
- Wetzel, S. A., McKeithan, T. W. & Parker, D. C. Peptide-specific intercellular transfer of MHC class II to CD4⁺ T cells directly from the immunological synapse upon cellular dissociation. *J. Immunol.* **174**, 80–89 (2005).
- Rosenits, K., Keppler, S. J., Vucikujia, S. & Aichele, P. T cells acquire cell surface determinants of APC via *in vivo* trogocytosis during viral infections. *Eur. J. Immunol.* **40**, 3450–3457 (2010).
- Daubeuf, S. et al. Preferential transfer of certain plasma membrane proteins onto T and B cells by trogocytosis. *PLoS One* **5**, e8716 (2010).
- Garside, P. et al. Visualization of specific B and T lymphocyte interactions in the lymph node. *Science* **281**, 96–99 (1998).
- Okada, T. et al. Antigen-engaged B cells undergo chemotaxis toward the T zone and form motile conjugates with helper T cells. *PLoS Biol.* **3**, e150 (2005).

Author contributions D.F.N. and J.B.W. designed the studies. R.B.J., R.A., E.C., Y.R., N.D.B., C.E.O., R.T. and S.Y.L. led the virology assays. S.H.H., D.C., J.L.K., M.A. and C.E.O. led the immunology assays. J.B.W. led the studies and wrote the paper with all co-authors.

Competing interests Declared none.

Additional information

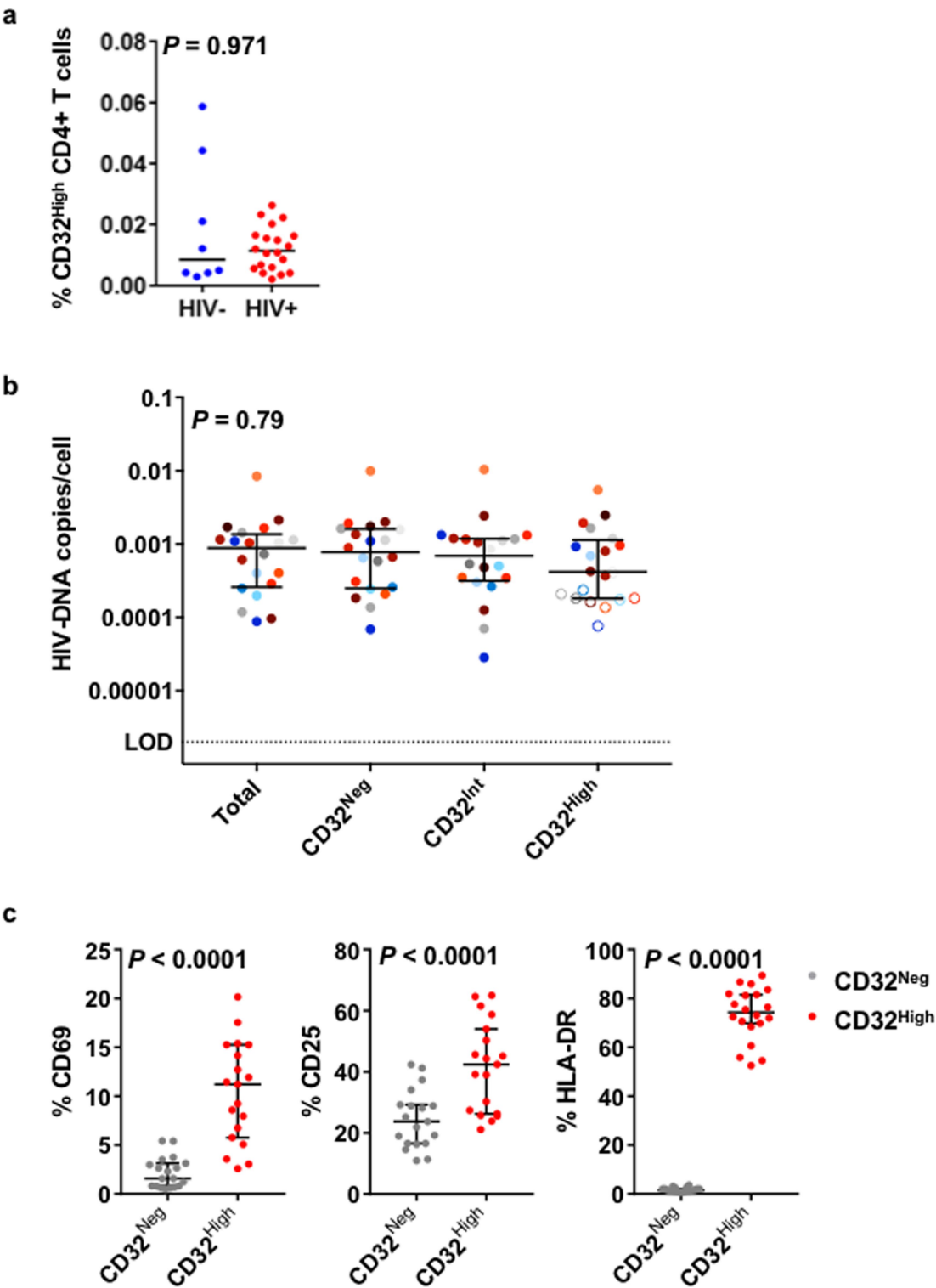
Extended data accompanies this Comment.

Supplementary information accompanies this Comment.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

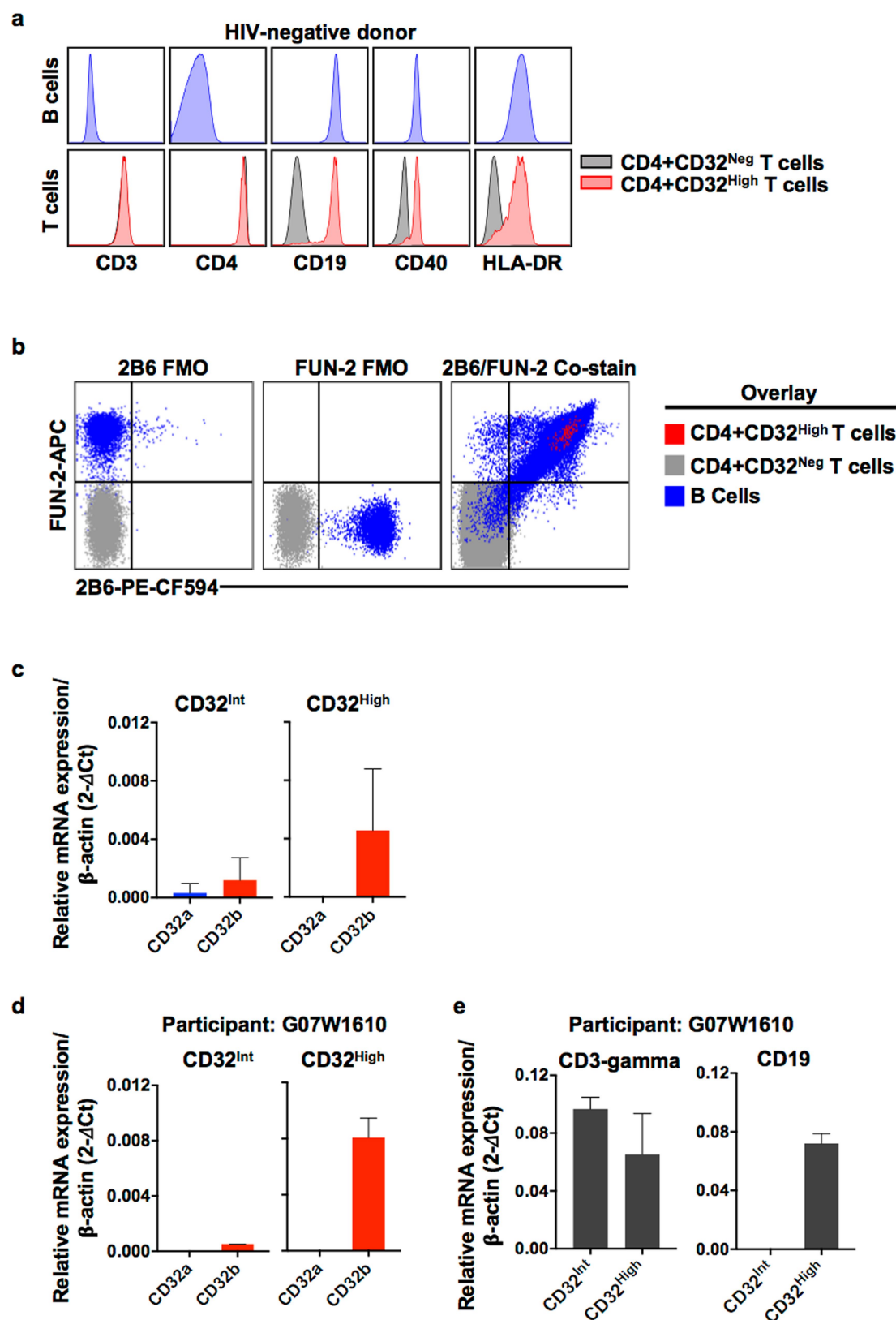
Correspondence and requests for materials should be addressed to J.B.W.

<https://doi.org/10.1038/s41586-018-0495-2>



Extended Data Fig. 1 | Frequency and activation status of CD32-expressing CD4⁺ T cells and their HIV-1 DNA content. **a**, The frequency of CD32^{high} CD4⁺ T cells was measured by flow cytometry in PBMCs from ART-suppressed, HIV-1⁺ ($n = 20$) and HIV-1⁻ ($n = 8$) donors. Bars denote median values. P values were determined by a Mann–Whitney test. **b**, DNA copies per cell in sorted subsets of total CD4⁺, CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells are shown with median values and the IQR. The results are shown as either the actual HIV-1 DNA

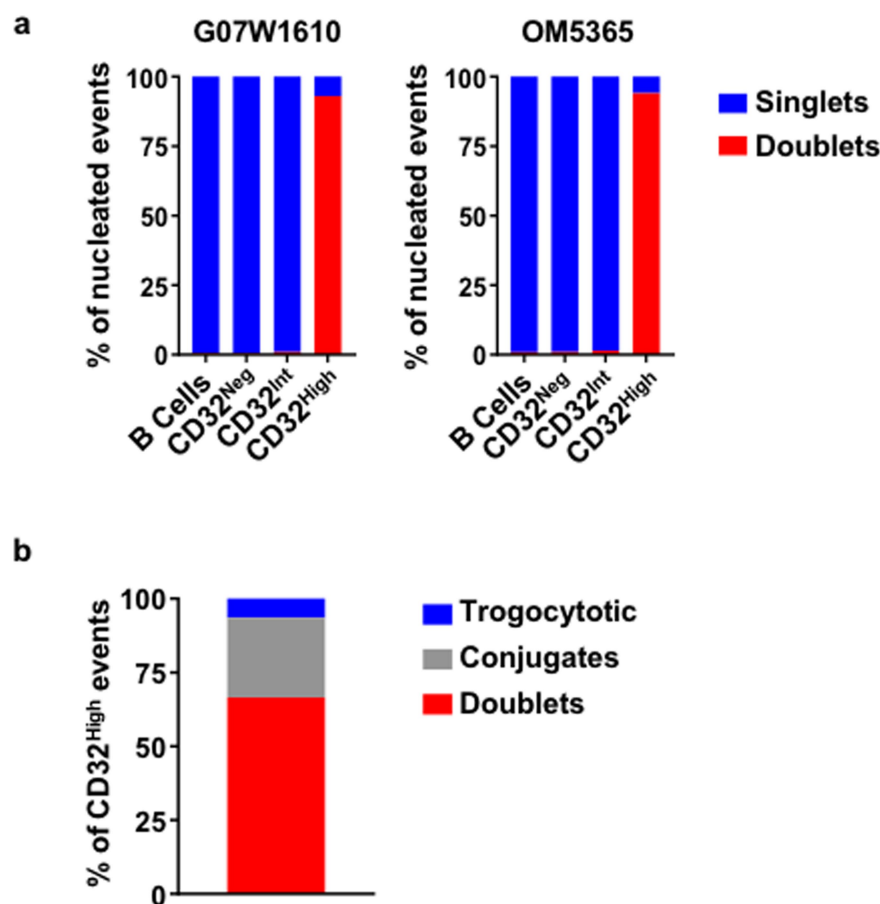
copies per million cells (filled symbols) or as estimated values calculated using the LOD and applied to the number of cells when the DNA input did not reach the threshold (open symbols). P values were determined by a Kruskal–Wallis test. **c**, The percentage of CD69, CD25 and HLA-DR expression was measured by flow cytometry on CD32^{neg} and CD32^{high} (FUN-2) CD4⁺ T cells from PBMCs from HIV-1⁺ participants ($n = 20$). Error bars show the median and IQR. P values were determined by Wilcoxon matched-pairs signed rank tests.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Detection of B cell proteins and mRNA in CD32-expressing CD4⁺ T cells. **a**, CD32^{neg} and CD32^{high} (FUN-2) CD4⁺ T cells from human PBMCs were assessed by flow cytometry for the expression of CD19, CD40 and HLA-DR, and compared to B cells (CD3⁻ CD14⁻ CD19⁺ lymphocytes). Representative flow cytometry results of per cell antigen levels on B cells (top, blue histograms) and CD32^{neg} and CD32^{high} CD4⁺ T cells (bottom, grey and red histograms, respectively) from an HIV-1⁻ donor. **b**, Representative CD32b staining of PBMCs from an HIV-1⁺, ART-suppressed participant. PBMCs were stained with an optimized concentration of the 2B6 monoclonal anti-CD32b antibody, followed by an antibody cocktail that included the FUN-2 monoclonal pan-CD32 antibody, as described in the Methods. Shown are the 2B6 and FUN-2

fluorescence minus one (FMO) antibody cocktail-stained samples and a sample co-stained with 2B6 and FUN-2. **c**, **d**, CD32 mRNA expression levels in CD4⁺CD32⁺ subsets. **c**, The relative expression of CD32A and CD32B mRNA isoforms in sorted CD4⁺CD32^{int} and CD4⁺CD32^{high} subsets from HIV-1⁺, ART-suppressed participants ($n = 4$). **d**, mRNA expression of CD32A and CD32B from patient G07W1610. **e**, T and B cell lineage-specific mRNA transcripts in sorted CD4⁺CD32⁺ subsets from participant G07W1610. Relative mRNA expression of target genes was normalized to *ATCB* using the comparative C_t method. Results are mean \pm s.d. of each value from each participant ($n = 4$; **c**), or from values generated from two separate experiments using samples from the same patient (**d**).



Extended Data Fig. 3 | Doublet composition of the sorted CD4⁺CD32^{high} T cells. Sorted B cells and CD4⁺CD32^{neg}, CD4⁺CD32^{int} and CD4⁺CD32^{high} T cells from an HIV-1⁺, ART-suppressed participant were analysed using an Amnis imaging cytometer. Singlets and doublets were quantified using the aspect ratio and nuclear staining. **a**, The proportion of total singlet and doublet events among total nucleated

cells detected on the Amnis cytometer in each sorted population was determined, and is shown as individual composite bar graphs for two patients (G07W1610 and OM5365). **b**, A composite bar graph of the proportion of conjugates, doublets and trogocytotic events that comprised the sorted CD4⁺CD32^{high} population ($n = 2$).

BRIEF COMMUNICATIONS ARISING

Extended Data Table 1 | Viral suppression of 20 HIV-1-infected participants on ART

Cohort	Participant ID	Date of Initial Suppression (MM/YY)	Length of suppression (yrs)
HEAL	HEAL-009	3/14	3
	HEAL-019	8/09	8
	HEAL-020	3/08	9.3
	HEAL-034	11/05	11.5
	HEAL-053	11/16	1
	HEAL-055	11/00	17
Maple Leaf	CIRC0024	6/98	17.0
	CIRC0133	7/08	7.0
	CIRC0196	4/14	1.2
	OM5011	11/08	6.6
	OM5148	1/08	7.5
	OM5162	9/04	10.8
	OM5203	3/12	3.3
	OM5334	7/14	0.9
	OM5365	3/08	7.3
WWH	WWH-B001	7/11	6.4
	WWH-B005	12/17	0.3
	WWH-B008	11/14	3.1
	WWH-B011	11/11	6
UPenn	G07W1610	10/05	11.8

BRIEF COMMUNICATIONS ARISING

Extended Data Table 2 | CD4⁺CD32^{high} subset proportions and HIV-1 DNA compared to total CD4⁺ and CD32^{neg} CD4⁺ T cells

Participant ID	CD32 ^{High}		HIV-DNA enrichment			
	% in total CD4	Absolute cell count	HIV-DNA copies/cell ¹	CD32 ^{High} /CD4 total ²	CD32 ^{High} /CD32 ^{Neg2}	CD32 ^{Neg} /CD4 total
HEAL-009	0.007	11,427	>0.000002	0.010	0.008	1.236
HEAL-019	0.002	4,911	>0.000002	0.002	0.001	1.500
HEAL-020	0.011	8,482	>0.000002	0.008	0.008	1.036
HEAL-034	0.022	8,238	0.000426	0.199	0.212	0.937
HEAL-053	0.004	9,544	>0.000002	0.017	0.015	1.161
HEAL-055	0.011	21,806	0.00037	0.604	0.555	1.088
CIRC0024	0.015	26,200	>0.000002	0.023	0.029	0.782
CIRC0133	0.017	14,602	>0.000002	0.005	0.010	0.517
CIRC0196	0.006	5,935	0.000694	1.722	1.066	1.615
OM5011	0.008	8,862	0.001942	1.871	2.187	0.855
OM5148	0.004	8,788	0.001191	1.043	1.049	0.994
OM5162	0.016	7,133	0.000923	0.842	0.842	1.000
OM5203	0.026	12,254	>0.000002	0.021	0.011	1.911
OM5334	0.016	6,275	0.000959	0.579	0.499	1.160
OM5365	0.006	11,027	>0.000002	0.003	0.003	0.803
WWH-B001	0.020	10,922	>0.000002	0.007	0.006	1.076
WWH-B005	0.013	5,964	0.00547	0.650	0.550	1.182
WWH-B008	0.023	6,464	0.000805	0.696	0.595	1.169
WWH-B011	0.004	5,953	0.001653	1.154	1.016	1.135
G07W1610	0.012	7,984	0.002482	1.452	1.411	1.029
Median	0.012	8,635	0.000398	0.389	0.356	1.082

¹Values below the LOD (2 copies per 10⁶ cells) are shaded in grey.

²To calculate HIV-1 enrichment, 0.000002 was used for all values below the LOD.

Descours et al. reply

REPLYING TO L. Pérez et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0493-4> (2018); C. E. Osuna et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0495-2> (2018); L. N. Bertagnolli et al. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0494-3> (2018)

In our previous work¹, we used an in vitro model of HIV-infected unstimulated CD4 T cells to identify CD32 as a candidate marker of HIV⁺ resting CD4 T cells in vitro, and a subset of HIV⁺ total CD4 T cells containing replication-competent viruses in individuals that underwent anti-retroviral therapy (ART). Of note, we did not explore the transcriptional status of hosted viruses (latent or active) ex vivo, nor the activation state of these cells (quiescent or activated)¹. In the accompanying Comments^{2–4}, colleagues attempted to reproduce these findings. They present experiments that support the following conclusions: (1) the isolation of the CD32⁺ CD4 T cell population results from artefacts caused by the flow cytometry sorting method^{2,3}, and (2) the sorted CD32 CD4 T cell population is not enriched in HIV nor in replication-competent proviral DNA^{2–4}. Here, we formulate two questions that mirror the major issues raised by these three Comments^{2–4} and discuss their results in the context of our previous report¹ and more recently published studies.

Is there any evidence that a CD4 T cell can express CD32 in the context of HIV infection? This question is raised by both Osuna et al.² and Pérez et al.³. A recent report⁷, using in situ hybridization (which avoids the criticism of artefacts caused by flow cytometry sorting), showed that HIV-1 RNA co-localized with CD32A (also known as FCGR2A) RNA in 90% of examined cells in B cell follicles from four individuals. Because HIV primarily targets CD4 T cells, these data may support the ability of a CD4 T cell to upregulate CD32 mRNA transcription after infection in vivo. Three independent groups have identified CD32 as being expressed by latently or productively infected CD4 T cells in vitro^{1,5–7}. These models generated and analysed a substantial percentage of HIV-infected CD4 cells. Thus, any marker that is usually not expressed by CD4 T cells but that is detected at the surface of these cells after infection is unlikely to result from biased analyses of cellular doublets, as could be the case when working on rare events from ex vivo samples^{2,3}. Instead, these data suggest that transcriptional regulation leading to the expression of CD32 mRNA and protein can probably occur after in vitro and in vivo infection of a single CD4 T cell.

Does the CD32 CD4 T cell subset contribute to viral persistence under treatment? All three of the accompanying Comments^{2–4} indicate that CD32 CD4 T cells are not enriched for HIV DNA in blood. Recent work suggests, however, that in some virally suppressed HIV-infected individuals, CD32 CD4 T cells were enriched in HIV DNA, although to a lesser extent than we reported⁸. Notably, this question has been recently addressed in tissues, and results seem to be less contrasted than in blood^{7,9,10}. More importantly, they revealed functional properties of these reservoir cells that have not been previously explored^{7,9,10}. As discussed above, a recent report⁷ found that within the B cell follicles of virally suppressed HIV-infected individuals, most of the cells containing HIV RNA and persisting despite treatment were found to express CD32A RNA⁷. This result seems to be in line with other data¹⁰ that indicate that T follicular helper cells, primarily found in these territories, were enriched for HIV DNA and RNA when expressing CD32¹⁰, although at a lower extent than our previous findings¹. In non-lymphoid rectal tissue, CD4 T cells expressing CD32 were also enriched

for both HIV DNA and RNA⁹. Notably, the co-expression of CD32 and HIV RNA reported in these two publications^{9,10} suggests that CD32 marks transcriptionally active infected cells rather than latent cells. Together, these reports support the ability of CD32 to identify a subset of persistent HIV-infected CD4 T cells and suggest that they could contribute to viral persistence under ART in vivo.

In conclusion, we believe that rather than completely ruling out the relevance of CD32 for the identification of a subset of infected cells in vivo and their contribution to HIV persistence, the whole literature, including the three accompanying Comments^{2–4}, opens new technical challenges and questions that we should solve in the near future.

Benjamin Descours, Gael Petitjean and Monsef Benkirane are solely responsible for this Reply. The contributions of the remaining authors from the original Letter¹ were limited to recruiting patients or performing analysis on blinded samples, and thus only Descours, Petitjean and Benkirane have authored this Reply.

Benjamin Descours¹, Gael Petitjean¹ & Monsef Benkirane^{1*}

¹Institut de Génétique Humaine, Laboratoire de Virologie Moléculaire, UMR9002, CNRS, Université de Montpellier, Montpellier, France.

*e-mail: monsef.benkirane@igh.cnrs.fr

1. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
2. Osuna, C. E. et al. Evidence that CD32a does not mark the HIV-1 latent reservoir. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0495-2> (2018).
3. Pérez, L. et al. Conflicting evidence for HIV enrichment in CD32⁺ CD4 T cells. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0493-4> (2018).
4. Bertagnolli, L. N. The role of CD32 during HIV-1 infection. *Nature* **561**, <https://doi.org/10.1038/s41586-018-0494-3> (2018).
5. Iglesias-Ussel, M., Vandergeeten, C., Marchionni, L., Chomont, N. & Romero, F. High levels of CD2 expression identify HIV-1 latently infected resting memory CD4⁺ T cells in virally suppressed subjects. *J. Virol.* **87**, 9148–9158 (2013).
6. Grau-Expósito, J. et al. A Novel single-cell FISH-flow assay identifies effector memory CD4⁺ T cells as a major niche for HIV-1 transcription in HIV-infected patients. *MBio* **8**, e00876-17 (2017).
7. Abdel-Mohsen, M. et al. CD32 is expressed on cells with transcriptionally active HIV but does not enrich for HIV DNA in resting T cells. *Sci. Transl. Med.* **10**, eaar6759 (2018).
8. Martin, G. E. et al. CD32-expressing CD4 T cells are phenotypically diverse and can contain proviral HIV DNA. *Front. Immunol.* **9**, 928 (2018).
9. Hogan, L. E. et al. Increased HIV- transcriptional activity and infectious burden in peripheral blood and gut-associated CD4⁺ T cells expressing CD30. *PLoS Pathog.* **4**, e006856 (2018).
10. Noto, A., Procopio, F., Corpataux, J. M. & Pantaleo, G. CD32⁺PD1⁺ Tfh cells are the major HIV reservoir in long-term art-treated individuals. *J. Virol.* <https://doi.org/10.1128/JVI.00901-18> (2018).

Author contributions B.D., G.P. and M.B. wrote the manuscript.

Competing interests Declared none.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.B.

<https://doi.org/10.1038/s41586-018-0496-1>

nature INDEX 2018 RISING STARS

NATURE, VOL. 561, ISSUE NO. 7723 (20 SEPTEMBER 2018)

THE FAST TRACK

If ideas are the flames burning from the torches of discovery, scientists are the hands that hold them. Creative minds uphold the scientific enterprise.

In recognition of their leading role, Nature Index 2018 Rising Stars profiles 11 up-and-coming researchers in the natural sciences (S10). These scientists are highlighted based on their recent contributions to the 82 journals tracked by the Nature Index, and their standing in the League of Scholars Whole-of-Web ranking, which assesses individuals on their research quality and impact, industry links and co-authorship networks. Their work ranges from analysing peatland and permafrost, to developing wearable electronics.

The researchers have all demonstrated excellence, and the passion, ambition and resilience to rise higher — essential for surviving in academia. As competition for jobs intensifies, researchers are expected to do more earlier in their careers, from publishing high-quality research to achieving impact, attracting funding, teaching, and cultivating international connections.

This supplement also tells the stories of institutions (S26), countries and regions (S20) that have exceeded expectations over the past three years in their contribution to the Nature Index. The ones we have selected as rising stars experienced exceptional absolute and percentage growth in their output of high-quality research, either across the breadth of subjects in the natural sciences, or in specific areas. As always in the Nature Index, our primary quantitative measurement is fractional count (FC) — a metric that accounts for the relative contribution of each author to an article. All FC figures are adjusted to 2017 levels.

A section on young universities explores their progress in the research world, achieved without the years of experience of established competitors (S30). We expect to see more of these high-flying contributors to the Nature Index, in reports of scientific inquiry and the testimonies of social change.

Smriti Mallapaty
Senior editor, Nature Index

The world at their feet



S10 The newcomers making their mark in science across the disciplines.

Challenger states

S20: Strength in different sectors, subjects, and regions contributes to national success.

Discovery relies on strong support staff

S24: A lack of trained administrators is holding African scientists back.

Movers and shakers

S26: The most improved institutions in the Nature Index 2015–2017.

Green shoots

S30: The young universities

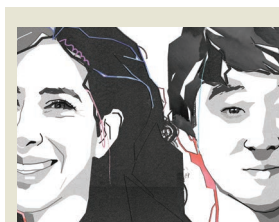
whose growth in high-quality research surpasses their peers.

Predicting scientific success

S32: Even sophisticated, data-driven models of academic careers have trouble forecasting the highs and lows.

The tables

S35: The world's institutions ranked by output rise.



ON THE COVER

Cognitive neuroscientist, Sarah Garfinkel, (left) and system engineer, Jaemin Kim, are bringing fresh ideas to their fields.

EDITORIAL: Catherine Armitage, Smriti Mallapaty, Rebecca Dargie, Herb Brody, Stephen Pincock **ANALYSIS:** Aaron Ballagh, Bo Wu, Willem Sijp **ART & DESIGN:** Tanner Maxwell, Madeline Hutchinson, Ruffi Lu, Mohamed Ashour, Wojtek Urbanek **WEB DEVELOPMENT & DESIGN:** Bob Edenbach, Olivier Lechevalier, Roberto Espinoza, Naomi Nakahara, Erika Suzuki **DATA QUALITY:** Jörn Ishikawa, Yuxin Wang, Paul Glaeser, Megha Katyal, Tipsy Jerald, Fabiha Shaikh, Rasika Zilpelwar **PRODUCTION:** Kay Lewis, Karl Smart, Ian Pope, Nick Bruni **MARKETING:** Stacy Best Ruel, Angelica Sarne **PROJECT MANAGEMENT:** Rebecca Jones, Chris Gilloch, Kazuki Kurebayashi, Sharon Wang **SALES:** Helen Hill, Janet Cen, Olaitan Fakindele, Maki Ishikawa, Neil Macmillan, Nichole Yu, Stella Yan, Tommy Yim **ART DIRECTOR:** Kelly Buckheit Krause **PUBLISHING:** Nick Campbell, Richard Hughes, David Swinbanks.

NATURE INDEX 2018 RISING STARS

Nature Index 2018 Rising Stars, a supplement to *Nature*, is produced by Nature Research, the flagship science portfolio of Springer Nature. This publication is based on data from the Nature Index, a Nature Research website maintained and made freely available at natureindex.com.

NATURE EDITORIAL OFFICES

The Campus, 4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

CUSTOMER SERVICES

To advertise with the Nature Index, please visit natureindex.com/client-services-feedback@nature.com
Copyright © 2018 Macmillan Publishers Limited, part of Springer Nature.
All rights reserved.

FORUM Structural biology

Views of light-activated proteins

The structures of anion-conducting channelrhodopsin proteins have been solved and used to develop a tool for optogenetics. Experts discuss what the structures tell us about ion conduction, and why the tool is needed. [SEE ARTICLES P.343 & 349](#)

THE PAPERS IN BRIEF

- Proteins called channelrhodopsins form light-activated ion channels in cell membranes.
- Channelrhodopsins have been used for optogenetics — a revolutionary technique that uses light to induce ion flux through these channels in genetically engineered cells, and thereby controls physiological processes.
- So far, the channelrhodopsins used most widely for optogenetics conduct positively charged ions (cations).
- However, the use of channelrhodopsins

that conduct negatively charged ions (anions) would open up new possibilities for optogenetics.

- In this issue, Kim *et al.*¹ (page 343) and Kato *et al.*² (page 349) describe the crystal structures of two anion-conducting channelrhodopsins, revealing the molecular basis of light-gated anion conduction, and providing insight into how these proteins could be modified for optogenetics.
- Kato *et al.* also report an engineered anion-conducting channelrhodopsin suitable for use in optogenetics.

For comparison, in CCRs, the intracellular region of the protein is mainly obstructed owing to the closure of the channel.

Kim *et al.* and Kato *et al.* report further analyses of *GtACR1* and *iC++* that provide insight into other aspects of how the structure of anion-conducting channelrhodopsins affects their function. For example, in *GtACR1*, they find that residues along the central constriction, in the retinal-binding pocket and at an extracellular region of the protein greatly affect the kinetics of pore closing. The two papers, in combination with two other recently described crystal structures of natural channelrhodopsins^{11,12}, provide a deeper structural and functional understanding of the light-activated ion-gating mechanism in microbial channelrhodopsins — and provide a basis for designing new classes of cation- and anion-conducting ion channels for optogenetics. ■

Structural insight

PATRICK SCHEERER

In the early 2000s, two channelrhodopsins were discovered^{3,4} in the microbial alga *Chlamydomonas reinhardtii*. Both are transmembrane receptors that have seven membrane-spanning α -helices and contain a deeply embedded molecule called retinal, which is covalently attached to the proteins, and is responsible for their light sensitivity. The proteins form channels that, when activated by light, allow various cations to flow down the electrochemical gradients that form across cell membranes when there are unequal concentrations of ions inside and outside the cell.

In 2012, the first crystal structure⁵ of a cation-conducting channelrhodopsin (CCR) was published — an engineered protein called C1C2, in which parts of the two *C. reinhardtii* channelrhodopsins were fused together. This breakthrough provided a snapshot of the structure of the light-activated channel and hinted at how channels are selectively conducted through it.

The C1C2 structure was subsequently used in conjunction with molecular modelling data to guide the design of engineered channelrhodopsins for use in optogenetics. As part of these efforts, several groups reported that certain structural modifications — including remodelling of the inner surface of the C1C2 pore, and specific mutations to the central part

of the protein that acts as the ion gate — could make channelrhodopsins anion-selective. This resulted in the development of highly chloride-selective channelrhodopsins^{6–9}, including a variant known as *iC++*, the structure of which is now reported by Kato and colleagues. A naturally occurring anion-conducting channelrhodopsin (*GtACR1*) was also discovered¹⁰ in the alga *Guillardia theta*, and its structure is reported by Kim and co-workers.

The structure of *GtACR1* shows that this protein shares a similar overall architecture with CCRs, but has several key differences (Fig. 1). For example, most of the amino acids at the surface of the *GtACR1* channel are positively charged (as is appropriate for cation exclusion), and not negatively charged as in CCRs. CCRs contain two extracellular vestibules (EV1 and EV2), only one of which (EV2) connects to the ion-conducting pathway. But in *GtACR1* (and also in *iC++*), both vestibules are remodelled, and only EV1 is connected to the ion-conducting pathway.

Although the structure of *GtACR1* captures the protein in its closed state, the channel is almost entirely open — remarkably, it is blocked by only one central constriction and at an extracellular constriction site in EV2. Kim *et al.* find that the central constriction contains three key amino-acid residues, two of which are involved in the process of anion transport, whereas the other is important for anion selectivity. A clear pathway allows anions released from the centre of the pore to reach the cell interior (although such release would not occur in this closed conformation).

Patrick Scheerer is in the Protein X-ray Crystallography and Signal Transduction Group, Institute of Medical Physics and Biophysics, Charité-Universitätsmedizin Berlin, 10117 Berlin, Germany.
e-mail: patrick.scheerer@charite.de

Tool development

ELIZABETH UNGER & LIN TIAN

Neurons in the brain function primarily by generating complex patterns of electrical impulses called action potentials, which were, for a long time, extremely cumbersome to measure and even more difficult to manipulate. That changed in 2005, when channelrhodopsin was introduced as a tool for optogenetics^{13,14}. When expressed in neurons, channelrhodopsin can be activated by light pulses to force those cells to fire action potentials. Optogenetics can precisely excite individual neurons or entire populations of genetically defined neurons in naturally behaving animals. This allowed direct testing of the contributions of different types of neuron to behavioural outcomes. The initial

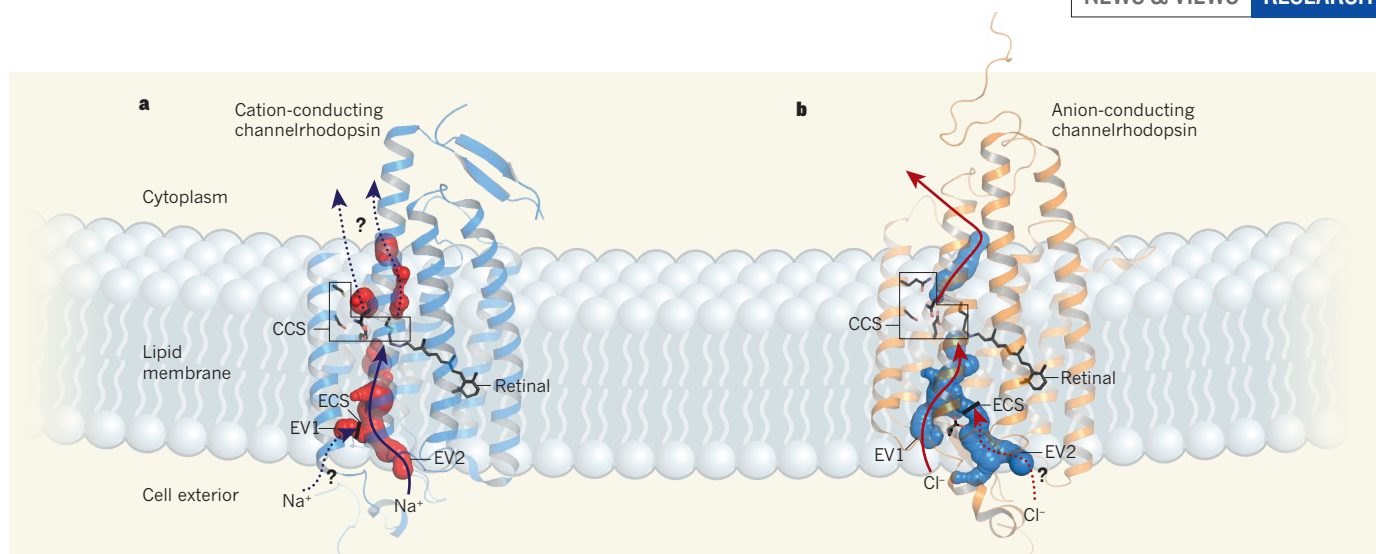


Figure 1 | The structures of cation-conducting and anion-conducting channelrhodopsin proteins. Channelrhodopsin proteins that transport positively charged ions (cations) such as sodium (Na^+) into cells have been widely used as tools for optogenetics — a technique that allows neurons to be activated by light pulses, aided by a light-sensitive molecule called retinal. Two papers^{1,2} now report the structures of channelrhodopsins that conduct negatively charged ions (anions) such as chloride (Cl^-) and the development of an engineered anion-conducting channelrhodopsin² that could be used in optogenetics to inhibit neuronal activity. **a, b.** The structure⁵ of a cation-conducting channelrhodopsin called C1C2 (**a**) and of the anion-conducting channelrhodopsin GtACR1 reported by Kim *et al.*¹ (**b**). Both proteins are shown in their closed conformation, in which the central ion-conducting channel is open, apart from a closed central constriction site (CCS). Some

of the key amino-acid residues of the CCS are shown, represented in a stick format. Solid arrows trace a possible route that ions might take through this central channel to enter the cell, and the path they would take out of GtACR1 if the channel was open. Dotted arrows indicate other ion-transit pathways that might be used and that would be blocked by an extracellular constriction site (ECS) or the closed CCS. The channel-exit path between the CCS and the cytoplasm is clearly visible as a continuous path only in the anion-conducting channel. The central cation-conducting channel (dark red) in **a** is mostly negatively charged, whereas the central anion-conducting channel (dark blue) in **b** is mostly positively charged. Another structural difference between the channels is that the extracellular-vestibule (EV) region that connects to the central ion-conducting pathway is EV2 in the cation-conducting channel and EV1 in the anion-conducting channel.

success of channelrhodopsin spawned an array of variants of the protein that have a range of useful properties for optogenetics¹⁵.

But just as much information is encoded by the absence of an action potential as by the presence of one. To fully understand the brain, we therefore need to be able to inhibit firing as well as to stimulate it — and to inhibit firing, proteins that move anions into cells in a manner that is controlled by light are needed. Unfortunately, optogenetic tools for inhibition have lagged considerably behind those for excitation. Light-activated proteins from the ion-pump family that could be used as inhibitory proteins were reported¹⁶ shortly after channelrhodopsin. But because the conductance of an ion pump is limited to one ion per absorbed photon of light, ion pumps never gained the same popularity.

In 2014, two groups independently reported^{6,8} genetically engineered chloride-conducting channelrhodopsins. Unlike ion pumps, these ion channels open to allow a large, rapid cellular influx of chloride anions in response to a single photon. These first-generation tools for optogenetics had only moderate sensitivity to light and poor temporal control, but were an excellent starting point from which improvements quickly followed.

Kato *et al.* now report the latest development: FLASH. This anion-conducting channelrhodopsin was designed using information gleaned from the two new crystal structures^{1,2}, and supplants the previous best-in-class protein for inhibitory optogenetics¹⁷, ZipACR. FLASH can suppress individual

action potentials in trains produced at frequencies of up to 40 hertz, with fewer off-target effects than ZipACR. It also inhibits neurons more reliably than ZipACR, reversibly reducing neuronal firing, on average, to about 30% of the initial rate in mouse brain slices. The authors show that when FLASH is expressed and activated in cells associated with the control of swimming in the nematode worm *Caenorhabditis elegans*, light almost completely inhibits swimming. We anticipate that any laboratory equipped for optogenetics should be able to use FLASH immediately.

But before FLASH can be widely adopted, proof is needed that its physiological side effects are minimal, and that it works in freely behaving fruit flies, zebrafish, mice and rats (four of the main animal models used by biologists). Moreover, temporal control of 40 Hz is probably acceptable for many applications, but 70% suppression of firing might not be sufficient for some experiments. And it remains to be seen whether long-term activation of the channel causes toxic effects in cells.

The current toolbox for inhibitory optogenetics is nowhere near as rich as that for excitatory experiments, but its expansion is following a similar timeline. We expect that the crystal structures reported by Kim *et al.* and Kato *et al.* will rapidly be used to engineer light-activated inhibitory channels that show substantial improvements over existing ones — FLASH represents a proof of concept of such efforts, rather than an end point. The new studies will surely inspire an array of tools to rival those of the excitatory

family, preferably, but not limited to, proteins that: totally suppress action potentials; are ultrasensitive to light and ultrafast; and are activated by different wavelengths of light. By using both inhibitory and excitatory manipulations, and combining these with the latest in genetic tools, imaging technology, behavioural assays and computational modelling, our understanding of how our powerful brains function will deepen considerably. ■

Elizabeth Unger and Lin Tian are in the Department of Biochemistry and Molecular Medicine, School of Medicine, University of California Davis, Davis, California 95616, USA. e-mails: lintian@ucdavis.edu; eunger@ucdavis.edu

- Kim, Y. S. *et al.* *Nature* **561**, 343–348 (2018).
- Kato, H. E. *et al.* *Nature* **561**, 349–354 (2018).
- Nagel, G. *et al.* *Science* **296**, 2395–2398 (2002).
- Nagel, G. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 13940–13945 (2003).
- Kato, H. E. *et al.* *Nature* **482**, 369–374 (2012).
- Berndt, A., Lee, S. Y., Ramakrishnan, C. & Deisseroth, K. *Science* **344**, 420–424 (2014).
- Berndt, A. *et al.* *Proc. Natl Acad. Sci. USA* **113**, 822–829 (2016).
- Wietek, J. *et al.* *Science* **344**, 409–412 (2014).
- Wietek, J. *et al.* *Sci. Rep.* **5**, 14807 (2015).
- Govorunova, E. G., Sineshchekov, O. A., Janz, R., Liu, X. & Spudich, J. L. *Science* **349**, 647–650 (2015).
- Volkov, O. *et al.* *Science* **358**, eaan8862 (2017).
- Li, H. *et al.* Preprint at bioRxiv <https://doi.org/10.1101/405308> (2018).
- Nagel, G. *et al.* *Curr. Biol.* **15**, 2279–2284 (2005).
- Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. *Nature Neurosci.* **8**, 1263–1268 (2005).
- Deisseroth, K. & Hegemann, P. *Science* **357**, eaan5544 (2017).
- Zhang, F. *et al.* *Nature* **446**, 633–639 (2007).
- Govorunova, E. G. *et al.* *Sci. Rep.* **7**, 43358 (2017).

ANIMAL MIGRATION

Bird forecasting

In our restless world, annual bird migrations can provide a transitory opportunity to glimpse beautiful avian visitors (pictured, the migratory Baltimore oriole; *Icterus galbula*). Birds' journeys are influenced by the local daily weather, making it hard to predict when migrating birds will pass through a particular place on their route. Writing in *Science*, Van Doren and Horton report a model that forecasts bird migrations (B. M. Van Doren and K. G. Horton *Science* **361**, 1115–1118; 2018).

They created their model (<http://birdcast.info>) by analysing bird migrations using 23 years of radar data from 143 locations across the United States, and assessing the data on weather conditions for the migrations. Being able to accurately predict an influx of birds might enable temporary measures to be taken to protect these migrants from hazards: for example, by turning off wind turbines. *Mary Abraham*



P. CHOU/GETTY

METROLOGY

Timing the action of light on matter

Photoemission, the ejection of an electron from a material on the absorption of a photon, is one of the fastest processes in nature. An experiment demonstrates how the dynamics of this process can be captured in real time. SEE LETTER P.374

THOMAS FENNEL

It seems natural that light facilitates photosynthesis, enables visual perception and provides the energy source for solar cells. But the underlying light-absorption process is not fully understood. Energy is transferred from the light to electrons in the irradiated material, which can cause electrons to be ejected — a phenomenon known as photoemission. The dependence of the electron ejection on the frequency of the incident light led to Albert Einstein's discovery¹ that light comes in discrete packets of energy (photons) and sparked the development of quantum mechanics. But how fast can an electron absorb a photon and escape? On page 374, Osslander *et al.*² show how metrology on the attosecond (10^{-18} seconds) timescale can help to answer this fundamental question.

It is only in the past decade or so that flashes of light could be generated that are short enough for researchers to directly track the dynamics of photoemission and to obtain timing information on the ejection of electrons^{3,4}. This advance has resulted in a vibrant revival of scientific interest in the fundamental physics of photoemission. The timing information contains valuable details about the electronic structure of the target material, many-body

effects (the correlated and collective behaviour of many interacting electrons) and the propagation of the electrons after photon absorption.

One of the key instruments used to carry out photoemission measurements is the attosecond streak camera⁵. In experiments based

on this instrument, a material is exposed to an attosecond-duration light pulse that has a frequency corresponding to the extreme-ultraviolet region of the electromagnetic spectrum. Electrons in the material absorb photons from the pulse and are ejected. These electrons are then accelerated by the electric field of a second light pulse — known as the streaking field — and the final energy of the electrons is measured.

Adjusting the time delay between the two light pulses changes the final electron energy in a well-defined way. This relationship enables a reconstruction of either the time evolution of the streaking field⁶ or the ejection time of the electrons^{3,4}, but not both simultaneously. As a result, streaking experiments have been unable to determine absolute photoemission delays — time differences between light absorption and electron ejection. Instead, they have provided

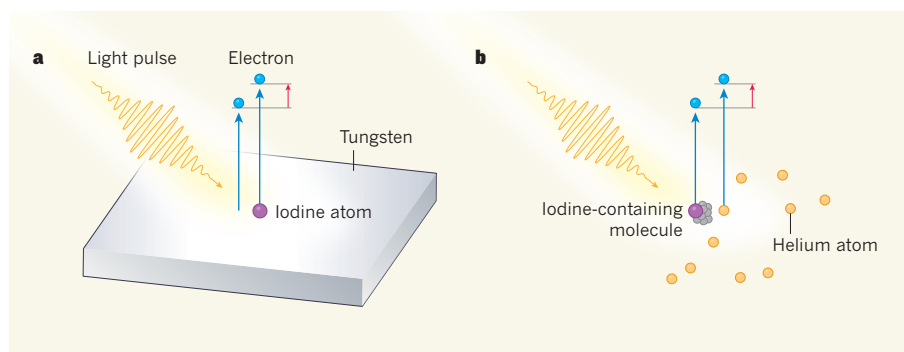


Figure 1 | Measurement of an absolute photoemission delay. Photoemission is the process by which an electron absorbs light and is ejected from a material. Osslander *et al.*² demonstrate a technique for determining absolute photoemission delays — time differences between light absorption and electron ejection. **a**, The authors deposited iodine molecules on a clean tungsten surface (for simplicity, a single iodine atom is shown here) and applied an extreme-ultraviolet light pulse to the material. Electrons were ejected (blue arrows) from both the tungsten surface and the iodine atoms. The authors measured the relative delay (red arrow) between these two ejections. **b**, Osslander and colleagues then applied the same light pulse to a gaseous mixture of iodine-containing molecules and helium atoms, and measured the relative iodine–helium photoemission delay. Finally, they used the known absolute photoemission delay for helium and the measured relative delays to determine the absolute photoemission delay for the tungsten surface.

ANIMAL MIGRATION

Bird forecasting

In our restless world, annual bird migrations can provide a transitory opportunity to glimpse beautiful avian visitors (pictured, the migratory Baltimore oriole; *Icterus galbula*). Birds' journeys are influenced by the local daily weather, making it hard to predict when migrating birds will pass through a particular place on their route. Writing in *Science*, Van Doren and Horton report a model that forecasts bird migrations (B. M. Van Doren and K. G. Horton *Science* **361**, 1115–1118; 2018).

They created their model (<http://birdcast.info>) by analysing bird migrations using 23 years of radar data from 143 locations across the United States, and assessing the data on weather conditions for the migrations. Being able to accurately predict an influx of birds might enable temporary measures to be taken to protect these migrants from hazards: for example, by turning off wind turbines. *Mary Abraham*



P. CHOU/GETTY

METROLOGY

Timing the action of light on matter

Photoemission, the ejection of an electron from a material on the absorption of a photon, is one of the fastest processes in nature. An experiment demonstrates how the dynamics of this process can be captured in real time. SEE LETTER P.374

THOMAS FENNEL

It seems natural that light facilitates photosynthesis, enables visual perception and provides the energy source for solar cells. But the underlying light-absorption process is not fully understood. Energy is transferred from the light to electrons in the irradiated material, which can cause electrons to be ejected — a phenomenon known as photoemission. The dependence of the electron ejection on the frequency of the incident light led to Albert Einstein's discovery¹ that light comes in discrete packets of energy (photons) and sparked the development of quantum mechanics. But how fast can an electron absorb a photon and escape? On page 374, Osslander *et al.*² show how metrology on the attosecond (10^{-18} seconds) timescale can help to answer this fundamental question.

It is only in the past decade or so that flashes of light could be generated that are short enough for researchers to directly track the dynamics of photoemission and to obtain timing information on the ejection of electrons^{3,4}. This advance has resulted in a vibrant revival of scientific interest in the fundamental physics of photoemission. The timing information contains valuable details about the electronic structure of the target material, many-body

effects (the correlated and collective behaviour of many interacting electrons) and the propagation of the electrons after photon absorption.

One of the key instruments used to carry out photoemission measurements is the attosecond streak camera⁵. In experiments based

on this instrument, a material is exposed to an attosecond-duration light pulse that has a frequency corresponding to the extreme-ultraviolet region of the electromagnetic spectrum. Electrons in the material absorb photons from the pulse and are ejected. These electrons are then accelerated by the electric field of a second light pulse — known as the streaking field — and the final energy of the electrons is measured.

Adjusting the time delay between the two light pulses changes the final electron energy in a well-defined way. This relationship enables a reconstruction of either the time evolution of the streaking field⁶ or the ejection time of the electrons^{3,4}, but not both simultaneously. As a result, streaking experiments have been unable to determine absolute photoemission delays — time differences between light absorption and electron ejection. Instead, they have provided

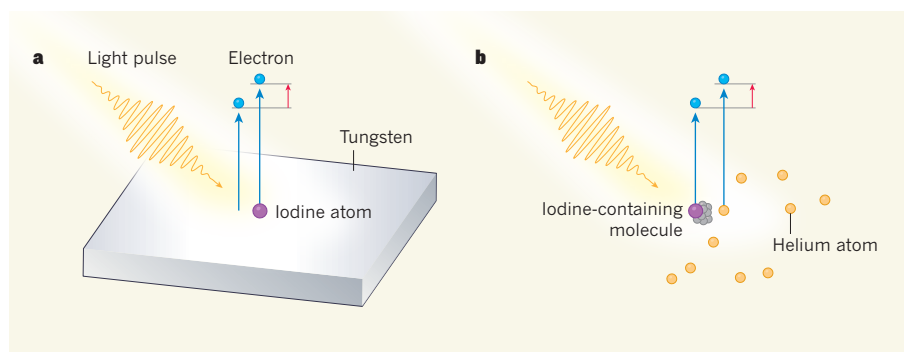


Figure 1 | Measurement of an absolute photoemission delay. Photoemission is the process by which an electron absorbs light and is ejected from a material. Osslander *et al.*² demonstrate a technique for determining absolute photoemission delays — time differences between light absorption and electron ejection. **a**, The authors deposited iodine molecules on a clean tungsten surface (for simplicity, a single iodine atom is shown here) and applied an extreme-ultraviolet light pulse to the material. Electrons were ejected (blue arrows) from both the tungsten surface and the iodine atoms. The authors measured the relative delay (red arrow) between these two ejections. **b**, Osslander and colleagues then applied the same light pulse to a gaseous mixture of iodine-containing molecules and helium atoms, and measured the relative iodine–helium photoemission delay. Finally, they used the known absolute photoemission delay for helium and the measured relative delays to determine the absolute photoemission delay for the tungsten surface.

measurements of relative delays, such as time differences between ejections of electrons from two different energy levels of the investigated material.

Ossiander and colleagues overcame this limitation using a clever two-stage approach, which they demonstrated by examining photoemission from a clean tungsten surface using an attosecond streak camera. In the first stage of the approach, the authors deposited iodine molecules on the tungsten surface (Fig. 1a). They then applied an attosecond-duration extreme-ultraviolet light pulse to the material and measured the relative delay in photoemission from the tungsten surface and from the atoms in the iodine molecules. In the second stage, the authors applied the same light pulse to a gaseous mixture of small iodine-containing molecules and helium atoms, and measured the relative iodine–helium photoemission delay (Fig. 1b).

Helium atoms are the largest atoms for which streaking experiments can currently be modelled completely by *ab initio* quantum simulations⁷. The absolute photoemission delay for helium is therefore known. Ossiander *et al.* used this result in combination with their measured relative delays to determine absolute photoemission delays for the tungsten surface. Their approach opens the door to measurements of such delays in surface and gas-phase experiments for many other target materials.

However, two central assumptions must be made when using Ossiander and colleagues' technique. First, additional delays caused by interactions between the iodine atoms and the target material must be negligible or known. Second, the iodine atoms must be close enough to the material's surface that spatial variations in the streaking field have only a small effect on the photoemission measurements. In the authors' experiment, the validity of these assumptions was backed up by theory. A closer analysis of the general limits in resolution associated with the technique will be a challenging, but important, task for future work.

The idea of using molecules as a reference to calibrate photoemission timing has previously been applied to streaking experiments on dielectric (insulating) nanoparticles⁸. These experiments suggest that photoemission delays could be used to directly characterize the attosecond-scale collisional dynamics of electrons in dielectric materials. The approach of Ossiander *et al.* is therefore expected to further advance the diagnostic capabilities of photoemission-delay measurements.

Ossiander *et al.* report photoemission delays for several different energy levels of the investigated tungsten surface. Their results imply that electron ejection from the material is more complex than was anticipated from previous measurements of relative delays³. The observed absolute delays can be explained only by considering both transport and collisional effects of the electrons during their propagation through the material.

A promising future application of the technique is the characterization of more-complex electronic effects — such as correlation, dissipation and decoherence — using data on absolute photoemission delays. This would provide a key reference for theory. The authors' observation of an extremely short delay (a few attoseconds) from the outermost electron shell of the iodine atoms also highlights application potential for ultrafast switching in electronic devices that operate at extremely high (petahertz; 10^{15} Hz) frequencies. Ossiander and colleagues have therefore provided insights into the dynamics of photoemission that not only advance our understanding of nature but also open routes to new technology. ■

BIOPHYSICS

Melting sculpts the embryo's body

Collections of cells in the tails of zebrafish embryos have now been found to transition between behaving as solids and fluids. This transition is responsible for the head-to-tail elongation of the embryo. [SEE LETTER P.401](#)

PIERRE-FRANÇOIS LENNE & VIKAS TRIVEDI

Understanding how different materials respond to force is central to the field of engineering. For instance, permanent application of force is required to deform a solid-like material, whereas a fluid-like material can be irreversibly deformed by transient forces. Over the past few decades, such concepts have also surfaced in biology. Much like inert materials such as foams and emulsions, collections of cells can switch from solid-like to fluid-like behaviours, depending on cell density and adherence. Processes that coordinate this tissue 'melting' with the application of forces have been shown to locally deform tissues while maintaining their global structure¹. Mongera *et al.*² report on page 401 that the elongation of the head-to-tail axis in zebrafish embryos relies on spatially controlled tissue 'melting'.

Head-to-tail (anterior-to-posterior) axis elongation is a central event in the generation of the animal body plan, and involves large-scale tissue deformation. For example, the posterior tip of a zebrafish embryo doubles in length in about five hours³. During this time, cells at the tip — in a region called the mesoderm progenitor zone (MPZ) — differentiate, becoming presomitic mesoderm (PSM) cells as they are left behind when posterior elongation proceeds. Cells of the PSM form structures called somites that will give rise to the animal's vertebrae (Fig. 1).

There are several known modes of tissue

Thomas Fennel is in the Theoretical Cluster Physics and Nanophotonics Group, Institute of Physics, University of Rostock, 18051 Rostock, Germany, and in the Attosecond Physics Division of the Max Born Institute for Nonlinear Optics and Short Pulse Spectroscopy, Berlin, Germany.
e-mail: thomas.fennel@uni-rostock.de

1. Einstein, A. *Ann. Phys.* **17**, 132–148 (1905).
2. Ossiander, M. *et al.* *Nature* **561**, 374–377 (2018).
3. Cavaliere, A. L. *et al.* *Nature* **449**, 1029–1032 (2007).
4. Schultze, M. *et al.* *Science* **328**, 1658–1662 (2010).
5. Itatani, J. *et al.* *Phys. Rev. Lett.* **88**, 173903 (2002).
6. Goulielmakis, E. *et al.* *Science* **305**, 1267–1269 (2004).
7. Ossiander, M. *et al.* *Nature Phys.* **13**, 280–285 (2017).
8. Seiffert, L. *et al.* *Nature Phys.* **13**, 766–770 (2017).

elongation. Polarized rearrangement of neighbouring cells can cause elongation in one direction and narrowing along a perpendicular axis⁴. In addition, external boundaries and forces can mediate elongation — neighbouring tissues can constrain, pull^{5,6} or compress^{7,8} tissues, and differences in the volume and stiffness of the extracellular matrix around cells can also provide guidance⁹. But, with a few exceptions¹⁰, we still do not know to what extent the material properties of cells as individuals and collectives control axis elongation *in vivo*, because it is technically challenging to simultaneously measure internal mechanical stresses and changing material properties within elongating tissues at cellular and supracellular scales.

Mongera *et al.* overcame this challenge by inserting magnetically responsive oil microdroplets between cells in the tails of zebrafish embryos undergoing elongation. They used changes in the shape of the microdroplets from spherical to ellipsoid to infer supracellular mechanical stresses, and so to map the spatial distribution of forces along the axis. First, the authors analysed the microdroplets in the absence of a magnetic field, which revealed a gradient of increasing force from the MPZ at the posterior tip of the embryo to the PSM. These supracellular stresses persisted for more than 30 minutes, on a par with the timescale over which PSM maturation leads to the formation of somites.

Second, the researchers applied a magnetic

measurements of relative delays, such as time differences between ejections of electrons from two different energy levels of the investigated material.

Ossiander and colleagues overcame this limitation using a clever two-stage approach, which they demonstrated by examining photoemission from a clean tungsten surface using an attosecond streak camera. In the first stage of the approach, the authors deposited iodine molecules on the tungsten surface (Fig. 1a). They then applied an attosecond-duration extreme-ultraviolet light pulse to the material and measured the relative delay in photoemission from the tungsten surface and from the atoms in the iodine molecules. In the second stage, the authors applied the same light pulse to a gaseous mixture of small iodine-containing molecules and helium atoms, and measured the relative iodine–helium photoemission delay (Fig. 1b).

Helium atoms are the largest atoms for which streaking experiments can currently be modelled completely by *ab initio* quantum simulations⁷. The absolute photoemission delay for helium is therefore known. Ossiander *et al.* used this result in combination with their measured relative delays to determine absolute photoemission delays for the tungsten surface. Their approach opens the door to measurements of such delays in surface and gas-phase experiments for many other target materials.

However, two central assumptions must be made when using Ossiander and colleagues' technique. First, additional delays caused by interactions between the iodine atoms and the target material must be negligible or known. Second, the iodine atoms must be close enough to the material's surface that spatial variations in the streaking field have only a small effect on the photoemission measurements. In the authors' experiment, the validity of these assumptions was backed up by theory. A closer analysis of the general limits in resolution associated with the technique will be a challenging, but important, task for future work.

The idea of using molecules as a reference to calibrate photoemission timing has previously been applied to streaking experiments on dielectric (insulating) nanoparticles⁸. These experiments suggest that photoemission delays could be used to directly characterize the attosecond-scale collisional dynamics of electrons in dielectric materials. The approach of Ossiander *et al.* is therefore expected to further advance the diagnostic capabilities of photoemission-delay measurements.

Ossiander *et al.* report photoemission delays for several different energy levels of the investigated tungsten surface. Their results imply that electron ejection from the material is more complex than was anticipated from previous measurements of relative delays³. The observed absolute delays can be explained only by considering both transport and collisional effects of the electrons during their propagation through the material.

A promising future application of the technique is the characterization of more-complex electronic effects — such as correlation, dissipation and decoherence — using data on absolute photoemission delays. This would provide a key reference for theory. The authors' observation of an extremely short delay (a few attoseconds) from the outermost electron shell of the iodine atoms also highlights application potential for ultrafast switching in electronic devices that operate at extremely high (petahertz; 10^{15} Hz) frequencies. Ossiander and colleagues have therefore provided insights into the dynamics of photoemission that not only advance our understanding of nature but also open routes to new technology. ■

BIOPHYSICS

Melting sculpts the embryo's body

Collections of cells in the tails of zebrafish embryos have now been found to transition between behaving as solids and fluids. This transition is responsible for the head-to-tail elongation of the embryo. [SEE LETTER P.401](#)

PIERRE-FRANÇOIS LENNE & VIKAS TRIVEDI

Understanding how different materials respond to force is central to the field of engineering. For instance, permanent application of force is required to deform a solid-like material, whereas a fluid-like material can be irreversibly deformed by transient forces. Over the past few decades, such concepts have also surfaced in biology. Much like inert materials such as foams and emulsions, collections of cells can switch from solid-like to fluid-like behaviours, depending on cell density and adherence. Processes that coordinate this tissue 'melting' with the application of forces have been shown to locally deform tissues while maintaining their global structure¹. Mongera *et al.*² report on page 401 that the elongation of the head-to-tail axis in zebrafish embryos relies on spatially controlled tissue 'melting'.

Head-to-tail (anterior-to-posterior) axis elongation is a central event in the generation of the animal body plan, and involves large-scale tissue deformation. For example, the posterior tip of a zebrafish embryo doubles in length in about five hours³. During this time, cells at the tip — in a region called the mesoderm progenitor zone (MPZ) — differentiate, becoming presomitic mesoderm (PSM) cells as they are left behind when posterior elongation proceeds. Cells of the PSM form structures called somites that will give rise to the animal's vertebrae (Fig. 1).

There are several known modes of tissue

Thomas Fennel is in the Theoretical Cluster Physics and Nanophotonics Group, Institute of Physics, University of Rostock, 18051 Rostock, Germany, and in the Attosecond Physics Division of the Max Born Institute for Nonlinear Optics and Short Pulse Spectroscopy, Berlin, Germany.
e-mail: thomas.fennel@uni-rostock.de

1. Einstein, A. *Ann. Phys.* **17**, 132–148 (1905).
2. Ossiander, M. *et al.* *Nature* **561**, 374–377 (2018).
3. Cavaliere, A. L. *et al.* *Nature* **449**, 1029–1032 (2007).
4. Schultze, M. *et al.* *Science* **328**, 1658–1662 (2010).
5. Itatani, J. *et al.* *Phys. Rev. Lett.* **88**, 173903 (2002).
6. Goulielmakis, E. *et al.* *Science* **305**, 1267–1269 (2004).
7. Ossiander, M. *et al.* *Nature Phys.* **13**, 280–285 (2017).
8. Seiffert, L. *et al.* *Nature Phys.* **13**, 766–770 (2017).

elongation. Polarized rearrangement of neighbouring cells can cause elongation in one direction and narrowing along a perpendicular axis⁴. In addition, external boundaries and forces can mediate elongation — neighbouring tissues can constrain, pull^{5,6} or compress^{7,8} tissues, and differences in the volume and stiffness of the extracellular matrix around cells can also provide guidance⁹. But, with a few exceptions¹⁰, we still do not know to what extent the material properties of cells as individuals and collectives control axis elongation *in vivo*, because it is technically challenging to simultaneously measure internal mechanical stresses and changing material properties within elongating tissues at cellular and supracellular scales.

Mongera *et al.* overcame this challenge by inserting magnetically responsive oil microdroplets between cells in the tails of zebrafish embryos undergoing elongation. They used changes in the shape of the microdroplets from spherical to ellipsoid to infer supracellular mechanical stresses, and so to map the spatial distribution of forces along the axis. First, the authors analysed the microdroplets in the absence of a magnetic field, which revealed a gradient of increasing force from the MPZ at the posterior tip of the embryo to the PSM. These supracellular stresses persisted for more than 30 minutes, on a par with the timescale over which PSM maturation leads to the formation of somites.

Second, the researchers applied a magnetic

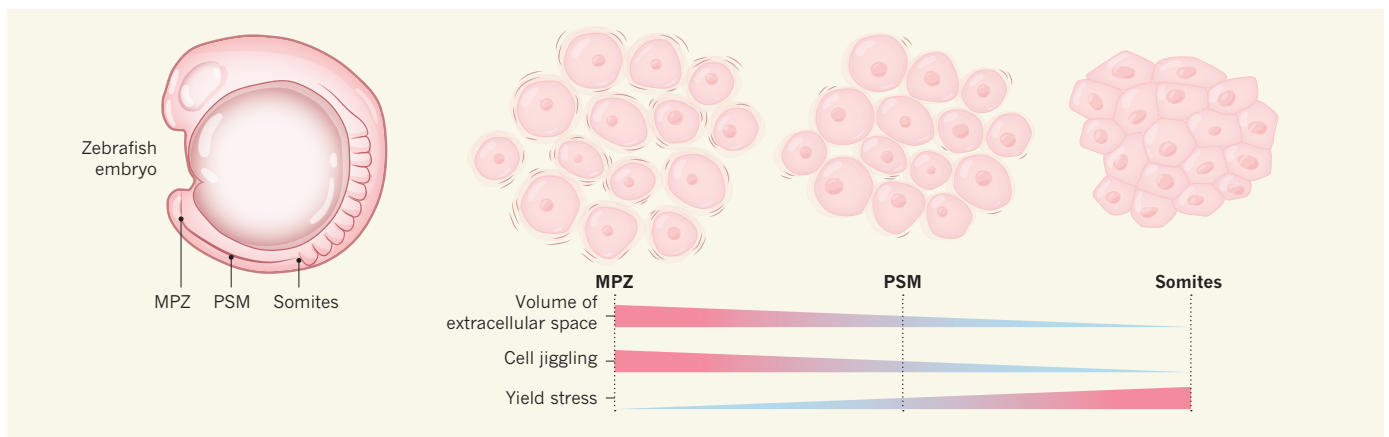


Figure 1 | A gradient of mechanical behaviour sculpts zebrafish.

Zebrafish embryos undergo a period of rapid elongation at their tails, during which cells originally in a mesodermal progenitor zone (MPZ) at the tip of the embryo become presomitic mesoderm (PSM) cells, and then form structures called somites that will give rise to vertebrae. Mongera *et al.*² have shown that cells in the MPZ behave in a fluid-like manner, and become progressively more solid-like as they mature into PSM and

somites. Changes in three physical factors govern this transition: decreases in the volume of extracellular space around cells; decreases in the rate at which cells ‘jiggle’, changing their contacts with neighbours; and increases in yield stress (the amount of stress needed to permanently deform the tissue). The gradients in yield stress and in the volume of extracellular space are controlled by the cell-adhesion protein N-cadherin (not shown), the concentration of which is similar in the MPZ and PSM.

field to the microdroplets to distend them, causing deformation of the tissue around them, and then investigated whether the droplets returned to their original spherical shape. This experiment revealed the amount of stress needed to permanently deform the tissue (a property called yield stress), which provides information about the material properties of the cells. Mongera *et al.* showed that yield stress also increases in a posterior-to-anterior direction, indicating that the MPZ is more fluid-like, and the PSM solid-like.

These measurements hint at the possibility of fluid-to solid ‘jamming’ — a concept well established to describe the transition of foam-like systems from wet to dry states¹¹. In foams, jamming depends on mechanical stresses, density and temperature^{11–13}. Mongera and colleagues hypothesized that equivalent parameters might govern the ability of the MPZ to ‘unjammed’ and behave in a fluid-like way. They therefore investigated, in addition to mechanical stresses, the volume of extracellular space between cells, and fluctuations in the contacts between cells (known as cell jiggling, a property often interpreted as the effective temperature) in their embryos.

This is a remarkable technical achievement, because measurements of all three parameters have not previously been made in the same system, even in inert soft materials. The experiments demonstrated that, whereas mechanical supracellular stresses decreased towards the MPZ, the volume of extracellular spaces increased, as did the extent of cell jiggling (Fig. 1). Of these factors, the researchers found that jiggling had the dominant role in keeping the MPZ unjammed. Their measurements of cell-scale mechanical stress (made by analysing small deformities in the ellipsoid nature of the microdroplets in the absence of a magnetic field) revealed that these stresses last only about one minute and show no

spatial bias along the anterior–posterior axis. However, because the yield stress is lower in the MPZ than in the PSM, cell-scale stress fluctuations are sufficient to drive cell jiggling in the MPZ and thereby tissue melting — by contrast, they fail to do so in the PSM.

Together, Mongera and colleagues’ data fit with typical scenarios for a jamming transition¹, in which the volume between interacting objects, here cells, is key to whether the objects behave as a fluid or a solid. In a final set of experiments, the authors show that the gradients in yield stress and in the volume of extracellular space are controlled by the cell-adhesion protein N-cadherin (although the concentration of the protein is not itself graded). The molecular mechanisms underpinning cell jiggling remain to be clarified, but the authors’

“Fluid-to-solid transitions are likely to occur in other animals, both in embryos and in adult organisms.”

work, in agreement with previous reports on 2D multicellular systems^{13,14}, show that the mechanics of cell–cell contacts — of adhesion in particular — have a prominent role in the jamming transition.

Fluid-to-solid transitions are likely to occur in other animals, both in embryos and in adult organisms, but we expect that the molecules that control them might differ from those that modulate axis elongation in zebrafish.

It will be interesting to determine how the parameters that control this transition in living materials relate to and differ from those at play in inert materials. In inert materials, the timescale over which material properties change is usually much larger than the timescale of typical deformations. By contrast, the mechanical properties of living matter can vary simultaneously with changes in the shape of tissues, and can, in turn, lead to further shape changes. This creates

a strong coupling between the overall material properties of the tissue, its shape changes as a result of cellular movements, and the force field generated and experienced by the constituent cells. We therefore expect living materials to provide us with a rich phenomenology, distinct from that of inert materials.

Axis elongation is widespread in development, and it will be important to look for hallmarks of similar transitions in other systems. It will be fascinating to learn more about how living systems obey the laws of physics using cellular and molecular strategies. ■

Pierre-François Lenne is at the Institut de Biologie du Développement de Marseille (IBDM), Aix Marseille University, CNRS, 13009 Marseille, France. **Vikas Trivedi** is at the European Molecular Biology Laboratory (EMBL) Barcelona, PRBB, 08003 Barcelona, Spain, and in the Department of Genetics, University of Cambridge, UK. e-mails: pierre-francois.lenne@univ-amu.fr; vikas.trivedi@embl.es

1. Park, J.-A. *et al.* *Nature Mater.* **14**, 1040–1048 (2015).
2. Mongera, A. *et al.* *Nature* **561**, 401–405 (2018).
3. Steventon, B. *et al.* *Development* **143**, 1732–1741 (2016).
4. Tada, M. & Heisenberg, C. P. *Development* **139**, 3897–3904 (2012).
5. Collinet, C. *et al.* *Nature Cell Biol.* **17**, 1247–1258 (2015).
6. Lye, C. M. *et al.* *PLoS Biol.* **13**, e1002292 (2015).
7. Bénazéraf, B. *et al.* *Nature* **466**, 248–252 (2010).
8. Jülich, D. *et al.* *Dev. Cell* **34**, 33–44 (2015).
9. Crest, J. *et al.* *eLife* **6**, e24958 (2017).
10. Clément, R. *et al.* *Curr. Biol.* **27**, 3132–3142 (2017).
11. Cohen-Addad, S., Höhler, R. & Pitois, O. *Annu. Rev. Fluid Mech.* **45**, 241–267 (2013).
12. van Hecke, M. J. *Phys. Condens. Matter* **22**, 033101 (2010).
13. Bi, D., Lopez, J. H., Schwarz, J. M. & Manning, M. L. *Nature Phys.* **11**, 1074–1079 (2015).
14. Farhadifar, R., Röper, J.-C., Aigouy, B., Eaton, S. & Jülicher, F. *Curr. Biol.* **17**, 2095–2104 (2007).

This article was published online on 5 September 2018.

SPINAL-CORD INJURY

Locomotion restored after paralysis

Spinal-cord injury can render intact neuronal circuits functionally dormant. Targeted reduction of neuronal inhibition in the injured region has now enabled reactivation of these circuits in mice, restoring basic locomotion.

GRÉGOIRE COURTINE

When we decide to walk, the brain broadcasts commands through parallel neuronal pathways that cascade to executive centres in the lumbar region of the spinal cord¹. A spinal-cord injury (SCI) scatters this exquisitely organized communication system, leading to severe locomotor deficits or paralysis². Most SCIs spare islands of intact neural tissues below the injury, which contain nerve fibres that remain connected to executive centres. But for unclear reasons, these anatomically intact connections remain functionally dormant. Writing in *Cell*, Chen *et al.*³ demonstrate that reducing the excitability of inhibitory neurons within the injured region of the spinal cord enables these dormant connections to relay commands from the brain, and promotes partial recovery of locomotion in mice that have sustained an SCI that causes complete paralysis.

It was long assumed that restoring movement after an SCI would involve precisely reconstituting the circuit connectivity that was in place before the injury. However, evidence of spontaneous circuit reorganization after SCI has challenged this view⁴. For example, consider injuries in which the spine is severed on either side in two staggered places. Although descending pathways from the brain are all severed at either the first or second lesion, new contacts can form between the projections from neurons in the brain that reach the second cut and local neurons that lie between the cuts. These contacts establish 'detour' circuits that relay sufficient information to executive centres to restore basic locomotion^{5,6} (Fig. 1). But until now, experimental procedures that trigger the formation and activation of these detour relays either have not been clinically relevant⁵ or have acted only transiently⁶.

Chen and colleagues set out to find a permanent way to render relay circuits functional following injury. They injected mice that had sustained a staggered SCI with a panel of compounds known to modulate neuronal activity. One of these compounds, a small molecule called CLP290, restored movement in the injured mice.

CLP290 activates a protein called KCC2 (ref. 7), which is a transporter of potassium and chloride ions, and is responsible for

maintaining a functional level of the latter in neurons. Neurotransmitter molecules such as GABA or glycine open chloride channels on the surface of target neurons, allowing chloride ions (Cl^-) to flow into the cells down a concentration gradient, resulting in neuronal inhibition. By pumping out Cl^- , KCC2 can control the concentration gradient and so limit how strongly target neurons are inhibited by these neurotransmitters.

SCIs lead to a decrease in the levels of KCC2 in neurons below the injury⁸. Chen *et al.* examined the effects of increasing KCC2 levels in their mice, either in the lumbar spinal cord below the SCIs or in the relay circuits between the staggered injuries. To do this, they used various genetically engineered mice, which enabled them to modulate KCC2 expression in the three main types of neuron in the spinal cord — inhibitory, excitatory and motor neurons. The authors found that increasing KCC2 expression in inhibitory neurons between the staggered lesions could reproduce the effects of CLP290 treatment, whereas no other manipulation could.

It would be intuitive to expect that downregulation of KCC2 following injury would

decrease the activity of neurons, by increasing the GABA- or glycine-mediated Cl^- influx. However, KCC2 downregulation actually leads to neuronal excitation if the lack of this transporter increases the concentration of Cl^- in the cells to such a level that GABA- or glycine-mediated opening of Cl^- channels causes an efflux, rather than an influx, of Cl^- (ref. 8). This mechanism enables GABA and glycine to excite neurons during development. After birth, KCC2 upregulation in neurons reduces intracellular Cl^- concentrations, transforming excitation into inhibition.

Similarly, the authors found that injury-mediated downregulation of KCC2 in inhibitory neurons located between the staggered lesions increased the cells' activity, and thus their ability to inhibit local relay circuits, rendering these circuits dormant. CLP290 restored the functionality of relay circuits by preventing downregulation of KCC2 and thus maintaining the balance between inhibition and excitation (Fig. 1). Moreover, the researchers showed that CLP290 did not affect the growth of new neuronal projections — recovery was triggered merely by restoring this balance in relay circuits.

Finally, Chen *et al.* demonstrated through two experiments that brain commands were being transmitted through the reactivated relays in CLP290-treated mice. First, electrically induced signals in the brain's cortex were relayed to the motor neurons below the injury, resulting in the activation of hindlimb muscles. Second, neurons between the staggered lesions were more active in response to locomotor activity in treated than in non-treated mice.

The authors' results are important. Together, they show that excessive neuron-mediated inhibition of relay circuits in the injured spinal cord is a key mechanism in preventing

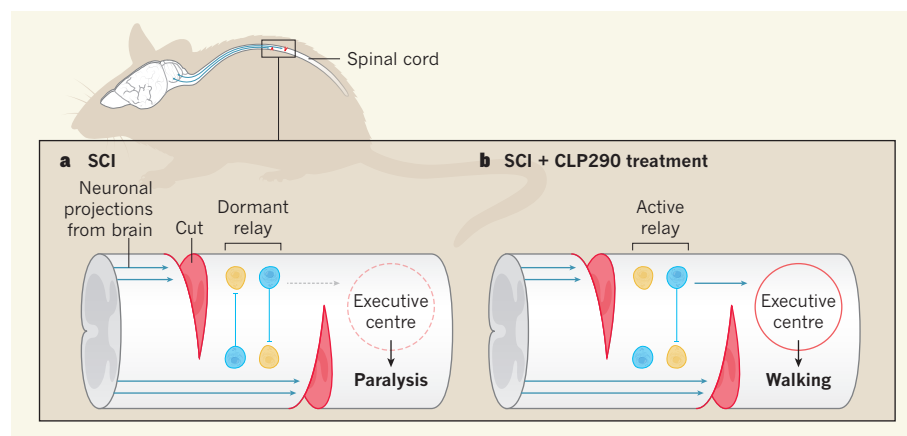


Figure 1 | Restoring balance in neuronal relay circuits. **a**, In a staggered spinal-cord injury (SCI), two partial cuts in the spine completely interrupt direct neuronal projections from the brain to executive centres that produce locomotion. Excitatory neurons (yellow) and inhibitory neurons (blue) located between the injuries can form relay circuits that, when active, pass brain-derived brain commands past the SCI. But Chen *et al.*³ show that, in mice with an SCI, excessive activity in inhibitory neurons increases inhibition of the circuits to a level that renders them dormant. Information is not passed to executive centres (indicated by dashed circle), and so the animals are paralysed. **b**, The authors find that treatment with a small molecule called CLP290 reduces the activity of the inhibitory neurons between the cuts. This restores the balance between excitation and inhibition, enabling relay circuits to pass information from the brain to executive centres and so leading to the recovery of basic locomotor movements such as walking.

anatomically intact but functionally dormant neuronal pathways from contributing to movement after an SCI.

What are the clinical implications of the findings? Chen *et al.* used a small molecule that is well tolerated in mice, even at a high concentration⁶. However, the relevance of this treatment for humans with severe SCI remains unclear. The authors' experimental model poorly mimics severe spinal-cord contusions commonly found in humans, in which nearly all the connections from putative relay neurons above the injured site are interrupted⁹. It therefore remains unclear whether Chen and colleagues' results could be reproduced after a clinically relevant SCI. Indeed, because CLP290 does not promote the growth of new neural connections, this treatment would be expected to be effective only after an SCI that spares a substantial proportion of nerve fibres. In addition, SCI causes a cascade of detrimental changes, so effective treatments must target multiple facets

of spinal-cord repair and recovery⁴, but CLP290 targets a single mechanism.

However, this type of orally administered pharmacological treatment is particularly attractive in combination with complementary strategies — notably, with interventions that promote the formation of relays in the spinal cord. For example, Chen and colleagues predict that CLP290 treatment will act synergistically with rehabilitative training, especially electrical spinal-cord stimulation, which promotes relay formation⁶. Alternatively, neural stem cells grafted into the injured spinal cord can enable reconstitution of relays across an SCI in monkeys¹⁰. Reducing neuron-mediated inhibition in the vicinity of the grafted relays could aid the functional integration of the relays into the host's neuronal networks.

We are reaching an exciting time in SCI medicine, when multiple interventions that have strong synergistic potential are approaching clinical applications. There are now

realistic opportunities to develop treatments that improve recovery after SCI in humans. ■

Grégoire Courtine is at the Center for Neuroprosthetics and the Brain Mind Institute, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland, and in the Department of Neurosurgery, Lausanne University Hospital. e-mail: gregoire.courtine@epfl.ch

1. Arber, S. & Costa, R. M. *Science* **360**, 1403–1404 (2018).
2. Fawcett, J. W. *et al. Spinal Cord* **45**, 190–205 (2007).
3. Chen, B. *et al. Cell* **174**, 521–535 (2018).
4. Sofroniew, M. V. *Nature* **557**, 343–350 (2018).
5. Courtine, G. *et al. Nature Med.* **14**, 69–74 (2008).
6. van den Brand, R. *et al. Science* **336**, 1182–1185 (2012).
7. Gagnon, M. *et al. Nature Med.* **19**, 1524–1528 (2013).
8. Boulenguez, P. *et al. Nature Med.* **16**, 302–307 (2010).
9. Asboth, L. *et al. Nature Neurosci.* **21**, 576–588 (2018).
10. Rosenzweig, E. S. *et al. Nature Med.* **24**, 484–490 (2018).

HIGH-ENERGY PHYSICS

Proton bunches rapidly accelerate electrons

Experiments show that short bunches of protons can produce electric fields that are strong enough to accelerate energetic electrons compactly. This discovery could lead to miniaturized high-energy particle accelerators. SEE LETTER P.363

TOSHIKI TAJIMA

For almost a century, particle accelerators have revealed the microscopic structure of the Universe in ever-increasing detail. This continual improvement has required progressively higher particle energies and, in turn, larger accelerators (the latest accelerator for such exploration¹ has a circumference of 27 kilometres). In conventional accelerators, particles are propelled by electromagnetic waves that are produced by external circuits. To drastically reduce the size of accelerators, scientists are exploring ways to use waves that

are instead generated internally, in an ionized gas known as a plasma². On page 363, Adli *et al.*³ report such a method, which makes use of an experiment in which the plasma waves are driven by bunches of protons — much like a motorboat on a lake drives waves in its wake.

The authors demonstrated their method using the Advanced Wakefield (AWAKE) experiment⁴, which is located at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. In this experiment, a proton bunch is injected into a plasma and sets electrons bobbing in its wake (Fig. 1). This electron motion generates a spatial modulation

in the electric-charge density of the plasma, which in turn produces an electric field known as a wakefield. If another electron is injected into the plasma a short distance behind the proton bunch, it is captured by the wakefield and is accelerated to high energies.

Because the proton bunch moves at close to the speed of light, the wakefield can be extremely strong. It can even be at the level of the Tajima–Dawson field², the amplitude of which is several orders of magnitude larger than that of the fields used in conventional accelerators. This is the reason that scientists see wakefield acceleration as a means of substantially miniaturizing particle accelerators.

The amplitude of a proton-driven wakefield can be so large only when the proton bunch and the plasma's internal clock (in this case, the oscillation period of the plasma waves) are in resonance — a condition that enhances the amplitude of the waves, akin to pushing a child on a swing synchronously with the swing's oscillation period. This condition is met when the length of the proton bunch matches the wavelength of the plasma waves. The plasma's ability to sustain strong fields increases when the plasma density is increased, which decreases the wavelength

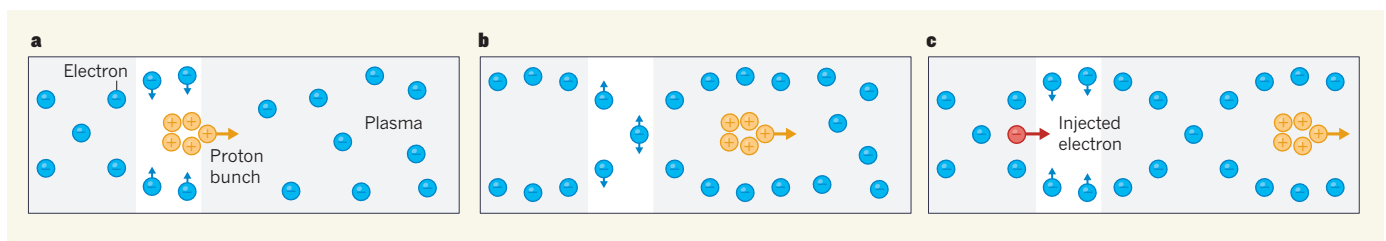


Figure 1 | The AWAKE experiment. **a**, In the Advanced Wakefield (AWAKE) experiment⁴, a bunch of protons is injected into an ionized gas known as a plasma. As the proton bunch travels through the plasma, it attracts electrons contained in the plasma, pulling them towards the centre. **b**, By the time these electrons have reached the centre, the proton bunch has moved

on. The electrons overshoot and begin to move outwards. **c**, The region that the electrons vacated is now positively charged. The electrons start to move inwards again, and the cycle repeats. Adli *et al.*³ show that if an electron is injected into the plasma a short distance behind the proton bunch, this cycling of positive and negative charge can rapidly accelerate the injected electron.

anatomically intact but functionally dormant neuronal pathways from contributing to movement after an SCI.

What are the clinical implications of the findings? Chen *et al.* used a small molecule that is well tolerated in mice, even at a high concentration⁶. However, the relevance of this treatment for humans with severe SCI remains unclear. The authors' experimental model poorly mimics severe spinal-cord contusions commonly found in humans, in which nearly all the connections from putative relay neurons above the injured site are interrupted⁹. It therefore remains unclear whether Chen and colleagues' results could be reproduced after a clinically relevant SCI. Indeed, because CLP290 does not promote the growth of new neural connections, this treatment would be expected to be effective only after an SCI that spares a substantial proportion of nerve fibres. In addition, SCI causes a cascade of detrimental changes, so effective treatments must target multiple facets

of spinal-cord repair and recovery⁴, but CLP290 targets a single mechanism.

However, this type of orally administered pharmacological treatment is particularly attractive in combination with complementary strategies — notably, with interventions that promote the formation of relays in the spinal cord. For example, Chen and colleagues predict that CLP290 treatment will act synergistically with rehabilitative training, especially electrical spinal-cord stimulation, which promotes relay formation⁶. Alternatively, neural stem cells grafted into the injured spinal cord can enable reconstitution of relays across an SCI in monkeys¹⁰. Reducing neuron-mediated inhibition in the vicinity of the grafted relays could aid the functional integration of the relays into the host's neuronal networks.

We are reaching an exciting time in SCI medicine, when multiple interventions that have strong synergistic potential are approaching clinical applications. There are now

realistic opportunities to develop treatments that improve recovery after SCI in humans. ■

Grégoire Courtine is at the Center for Neuroprosthetics and the Brain Mind Institute, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), 1015 Lausanne, Switzerland, and in the Department of Neurosurgery, Lausanne University Hospital. e-mail: gregoire.courtine@epfl.ch

1. Arber, S. & Costa, R. M. *Science* **360**, 1403–1404 (2018).
2. Fawcett, J. W. *et al. Spinal Cord* **45**, 190–205 (2007).
3. Chen, B. *et al. Cell* **174**, 521–535 (2018).
4. Sofroniew, M. V. *Nature* **557**, 343–350 (2018).
5. Courtine, G. *et al. Nature Med.* **14**, 69–74 (2008).
6. van den Brand, R. *et al. Science* **336**, 1182–1185 (2012).
7. Gagnon, M. *et al. Nature Med.* **19**, 1524–1528 (2013).
8. Boulenguez, P. *et al. Nature Med.* **16**, 302–307 (2010).
9. Asboth, L. *et al. Nature Neurosci.* **21**, 576–588 (2018).
10. Rosenzweig, E. S. *et al. Nature Med.* **24**, 484–490 (2018).

HIGH-ENERGY PHYSICS

Proton bunches rapidly accelerate electrons

Experiments show that short bunches of protons can produce electric fields that are strong enough to accelerate energetic electrons compactly. This discovery could lead to miniaturized high-energy particle accelerators. SEE LETTER P.363

TOSHIKI TAJIMA

For almost a century, particle accelerators have revealed the microscopic structure of the Universe in ever-increasing detail. This continual improvement has required progressively higher particle energies and, in turn, larger accelerators (the latest accelerator for such exploration¹ has a circumference of 27 kilometres). In conventional accelerators, particles are propelled by electromagnetic waves that are produced by external circuits. To drastically reduce the size of accelerators, scientists are exploring ways to use waves that

are instead generated internally, in an ionized gas known as a plasma². On page 363, Adli *et al.*³ report such a method, which makes use of an experiment in which the plasma waves are driven by bunches of protons — much like a motorboat on a lake drives waves in its wake.

The authors demonstrated their method using the Advanced Wakefield (AWAKE) experiment⁴, which is located at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. In this experiment, a proton bunch is injected into a plasma and sets electrons bobbing in its wake (Fig. 1). This electron motion generates a spatial modulation

in the electric-charge density of the plasma, which in turn produces an electric field known as a wakefield. If another electron is injected into the plasma a short distance behind the proton bunch, it is captured by the wakefield and is accelerated to high energies.

Because the proton bunch moves at close to the speed of light, the wakefield can be extremely strong. It can even be at the level of the Tajima–Dawson field², the amplitude of which is several orders of magnitude larger than that of the fields used in conventional accelerators. This is the reason that scientists see wakefield acceleration as a means of substantially miniaturizing particle accelerators.

The amplitude of a proton-driven wakefield can be so large only when the proton bunch and the plasma's internal clock (in this case, the oscillation period of the plasma waves) are in resonance — a condition that enhances the amplitude of the waves, akin to pushing a child on a swing synchronously with the swing's oscillation period. This condition is met when the length of the proton bunch matches the wavelength of the plasma waves. The plasma's ability to sustain strong fields increases when the plasma density is increased, which decreases the wavelength

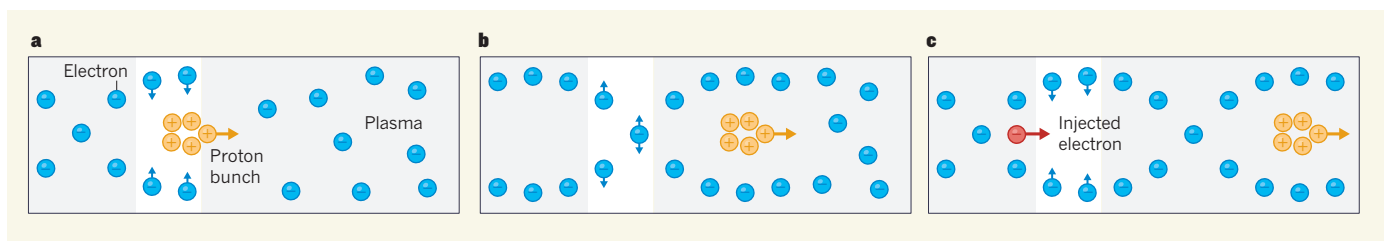


Figure 1 | The AWAKE experiment. **a**, In the Advanced Wakefield (AWAKE) experiment⁴, a bunch of protons is injected into an ionized gas known as a plasma. As the proton bunch travels through the plasma, it attracts electrons contained in the plasma, pulling them towards the centre. **b**, By the time these electrons have reached the centre, the proton bunch has moved

on. The electrons overshoot and begin to move outwards. **c**, The region that the electrons vacated is now positively charged. The electrons start to move inwards again, and the cycle repeats. Adli *et al.*³ show that if an electron is injected into the plasma a short distance behind the proton bunch, this cycling of positive and negative charge can rapidly accelerate the injected electron.

of the waves. Consequently, a stronger wakefield requires a shorter proton bunch.

The main innovation in Adli and colleagues' work was, therefore, to make the length of the proton bunch as short as possible so that the bunch resonates with the plasma's internal clock, maximizing the amplitude of the wakefield. The authors achieved this feat using a feature of the plasma known as collective force. Although the electric force produced by each particle in the plasma is small, the collective force generated from all of the particles can be large, and becomes larger as the plasma density is increased². The authors used this force to chop a long proton bunch into a series of shorter bunches. Because proton bunches are stiff (difficult to deform) at the extremely high particle energies present in the AWAKE experiment, this chopping was possible and effective only by using the plasma's collective force.

Adli *et al.* found that the wakefield produced by the short proton bunches could accelerate electrons to energies of up to 2 gigaelectronvolts in a plasma that is only about 10 metres in length. For comparison, at the European X-ray free-electron laser facility (European XFEL) in Germany, electrons are accelerated to energies of up to 17.5 gigaelectronvolts in an accelerator that is about 2 km long (see go.nature.com/2n6857t). In addition to providing compact acceleration, the authors' approach has a key advantage over standard accelerators and other wakefield accelerators. Because the proton bunches are stiff, they maintain their structure

and speed. As a result, high-energy electrons can be produced in a single acceleration stage, as opposed to the complex multi-stage process that is needed in other accelerators.

Usually, the higher the energy of a particle beam, the longer it takes to stop (dump) the beam after use. The dumping of high-energy beams has become a serious issue because of the requirement of longer dumping lengths, which in turn increases the production of unwanted radioactive isotopes in the dense materials used for the dumping. The authors show that their accelerated electrons can form a beam of short electron bunches, which would encounter a large collective force if injected into an appropriately prepared plasma. Such a beam could therefore be stopped over a much shorter distance than conventional beams, inducing little radioactivity⁵. Overall, the authors' work represents a major step towards the development of future high-energy particle accelerators that use collective force. ■

Toshiki Tajima is in the Department of Physics and Astronomy, University of California, Irvine, California 92697, USA.
e-mail: tajima@uci.edu

1. Evans, L. & Bryant, P. J. *Instrum* **3**, S08001 (2008).
2. Tajima, T. & Dawson, J. M. *Phys. Rev. Lett.* **43**, 267–270 (1979).
3. Adli, E. *et al.* *Nature* **561**, 363–367 (2018).
4. Gschwendtner, E. *et al.* *Nucl. Instrum. Meth. Phys. Res. A* **829**, 76–82 (2016).
5. Wu, H.-C., Tajima, T., Habs, D., Chao, A. W. & Meyer-ter-Vehn, J. *Phys. Rev. ST Accel. Beams* **13**, 101303 (2010).

CANCER

T cells home in on brain tumours

Immunotherapies activate T cells to destroy tumours, but the approach has failed in some brain cancers. A strategy to improve migration of T cells across the blood–brain barrier could overcome this limitation. [SEE ARTICLE P.331](#)

MICHAEL PLATTEN

Therapies that activate immune cells called T cells to target tumours are an efficient way to combat many types of cancer¹. But an aggressive brain cancer called glioblastoma has proved a particular challenge for immunotherapies². The blood–brain barrier protects the brain against immune-cell infiltration to prevent the potentially life-threatening effects of brain inflammation. This phenomenon is beneficial in normal circumstances, but it prevents T cells from reaching glioblastomas, making the tumours immunologically 'cold'³. On page 331, Samaha and colleagues⁴ report a way to trigger infiltration of T cells into the brains of mice, thus making

glioblastomas vulnerable to immunotherapy.

In the disease encephalitis, brain inflammation occurs because T cells that are typically excluded from the brain migrate across the blood–brain barrier. This migration is a coordinated process that requires activated T cells circulating in the bloodstream to adhere to endothelial cells, which line blood vessels. Adhesion is mediated by the binding of ligand molecules on T cells to cell-adhesion molecules such as ALCAM, ICAM-1 and VCAM-1 on endothelial cells⁵. These cell-adhesion molecules are expressed at higher than normal levels in encephalitis⁶. Binding between ALCAM and the T-cell ligand CD6 halts the progress of activated T cells through blood vessels, allowing subsequent binding by ICAM-1 and VCAM-1.



50 Years Ago

A campaign was opened last week for funds to refloat the Great Britain, one of the three major ships designed by Brunel. The object is to tow her back from the Falkland Islands to the Bristol shipyard ... The Great Britain was the first ocean-going iron ship and the first to be driven by propeller ... Brunel intended the ship to carry passengers of the Great Western Railway ... to New York, but the Great Britain made only a few transatlantic voyages before running aground ... Brunel managed to refloat the ship, which for the next 20 years carried emigrants to Australia ... In 1875, the Great Britain's engines were removed and she was converted to sail, plying between Liverpool and San Francisco until put out of service by a fire near the Falkland Islands ... Despite the ship's age, her structure is still sound enough to survive the journey back to Britain.
From *Nature* 21 September 1968

100 Years Ago

On the afternoon of Saturday, August 24 last, the allotment-holders of a small area in Hendon ... were sheltering in their sheds during a heavy thundershower, when they observed that small fish were being rained to the ground. The fish were precipitated on three adjoining roads and on the allotment-gardens enclosed by the roads; the rain swept them from the roads into the gutters and from the roofs of the sheds ... It is not easy to say how many fish fell, but ... they were numerous ... All the examples which came into my hands ... prove to be the lesser sand-eel (*Ammodytes tobianus*) ... The place where the sand-eels in question were deposited lies about one-quarter of a mile from the seashore ... The only explanation ... is that a shoal of sand-eels was drawn up by a waterspout.
From *Nature* 19 September 1918

of the waves. Consequently, a stronger wakefield requires a shorter proton bunch.

The main innovation in Adli and colleagues' work was, therefore, to make the length of the proton bunch as short as possible so that the bunch resonates with the plasma's internal clock, maximizing the amplitude of the wakefield. The authors achieved this feat using a feature of the plasma known as collective force. Although the electric force produced by each particle in the plasma is small, the collective force generated from all of the particles can be large, and becomes larger as the plasma density is increased². The authors used this force to chop a long proton bunch into a series of shorter bunches. Because proton bunches are stiff (difficult to deform) at the extremely high particle energies present in the AWAKE experiment, this chopping was possible and effective only by using the plasma's collective force.

Adli *et al.* found that the wakefield produced by the short proton bunches could accelerate electrons to energies of up to 2 gigaelectronvolts in a plasma that is only about 10 metres in length. For comparison, at the European X-ray free-electron laser facility (European XFEL) in Germany, electrons are accelerated to energies of up to 17.5 gigaelectronvolts in an accelerator that is about 2 km long (see go.nature.com/2n6857t). In addition to providing compact acceleration, the authors' approach has a key advantage over standard accelerators and other wakefield accelerators. Because the proton bunches are stiff, they maintain their structure

and speed. As a result, high-energy electrons can be produced in a single acceleration stage, as opposed to the complex multi-stage process that is needed in other accelerators.

Usually, the higher the energy of a particle beam, the longer it takes to stop (dump) the beam after use. The dumping of high-energy beams has become a serious issue because of the requirement of longer dumping lengths, which in turn increases the production of unwanted radioactive isotopes in the dense materials used for the dumping. The authors show that their accelerated electrons can form a beam of short electron bunches, which would encounter a large collective force if injected into an appropriately prepared plasma. Such a beam could therefore be stopped over a much shorter distance than conventional beams, inducing little radioactivity⁵. Overall, the authors' work represents a major step towards the development of future high-energy particle accelerators that use collective force. ■

Toshiki Tajima is in the Department of Physics and Astronomy, University of California, Irvine, California 92697, USA.
e-mail: tajima@uci.edu

1. Evans, L. & Bryant, P. J. *Instrum* **3**, S08001 (2008).
2. Tajima, T. & Dawson, J. M. *Phys. Rev. Lett.* **43**, 267–270 (1979).
3. Adli, E. *et al.* *Nature* **561**, 363–367 (2018).
4. Gschwendtner, E. *et al.* *Nucl. Instrum. Meth. Phys. Res. A* **829**, 76–82 (2016).
5. Wu, H.-C., Tajima, T., Habs, D., Chao, A. W. & Meyer-ter-Vehn, J. *Phys. Rev. ST Accel. Beams* **13**, 101303 (2010).

CANCER

T cells home in on brain tumours

Immunotherapies activate T cells to destroy tumours, but the approach has failed in some brain cancers. A strategy to improve migration of T cells across the blood–brain barrier could overcome this limitation. [SEE ARTICLE P.331](#)

MICHAEL PLATTEN

Therapies that activate immune cells called T cells to target tumours are an efficient way to combat many types of cancer¹. But an aggressive brain cancer called glioblastoma has proved a particular challenge for immunotherapies². The blood–brain barrier protects the brain against immune-cell infiltration to prevent the potentially life-threatening effects of brain inflammation. This phenomenon is beneficial in normal circumstances, but it prevents T cells from reaching glioblastomas, making the tumours immunologically 'cold'³. On page 331, Samaha and colleagues⁴ report a way to trigger infiltration of T cells into the brains of mice, thus making

glioblastomas vulnerable to immunotherapy.

In the disease encephalitis, brain inflammation occurs because T cells that are typically excluded from the brain migrate across the blood–brain barrier. This migration is a coordinated process that requires activated T cells circulating in the bloodstream to adhere to endothelial cells, which line blood vessels. Adhesion is mediated by the binding of ligand molecules on T cells to cell-adhesion molecules such as ALCAM, ICAM-1 and VCAM-1 on endothelial cells⁵. These cell-adhesion molecules are expressed at higher than normal levels in encephalitis⁶. Binding between ALCAM and the T-cell ligand CD6 halts the progress of activated T cells through blood vessels, allowing subsequent binding by ICAM-1 and VCAM-1.



50 Years Ago

A campaign was opened last week for funds to refloat the Great Britain, one of the three major ships designed by Brunel. The object is to tow her back from the Falkland Islands to the Bristol shipyard ... The Great Britain was the first ocean-going iron ship and the first to be driven by propeller ... Brunel intended the ship to carry passengers of the Great Western Railway ... to New York, but the Great Britain made only a few transatlantic voyages before running aground ... Brunel managed to refloat the ship, which for the next 20 years carried emigrants to Australia ... In 1875, the Great Britain's engines were removed and she was converted to sail, plying between Liverpool and San Francisco until put out of service by a fire near the Falkland Islands ... Despite the ship's age, her structure is still sound enough to survive the journey back to Britain.
From *Nature* 21 September 1968

100 Years Ago

On the afternoon of Saturday, August 24 last, the allotment-holders of a small area in Hendon ... were sheltering in their sheds during a heavy thundershower, when they observed that small fish were being rained to the ground. The fish were precipitated on three adjoining roads and on the allotment-gardens enclosed by the roads; the rain swept them from the roads into the gutters and from the roofs of the sheds ... It is not easy to say how many fish fell, but ... they were numerous ... All the examples which came into my hands ... prove to be the lesser sand-eel (*Ammodytes tobianus*) ... The place where the sand-eels in question were deposited lies about one-quarter of a mile from the seashore ... The only explanation ... is that a shoal of sand-eels was drawn up by a waterspout.
From *Nature* 19 September 1918

of the waves. Consequently, a stronger wakefield requires a shorter proton bunch.

The main innovation in Adli and colleagues' work was, therefore, to make the length of the proton bunch as short as possible so that the bunch resonates with the plasma's internal clock, maximizing the amplitude of the wakefield. The authors achieved this feat using a feature of the plasma known as collective force. Although the electric force produced by each particle in the plasma is small, the collective force generated from all of the particles can be large, and becomes larger as the plasma density is increased². The authors used this force to chop a long proton bunch into a series of shorter bunches. Because proton bunches are stiff (difficult to deform) at the extremely high particle energies present in the AWAKE experiment, this chopping was possible and effective only by using the plasma's collective force.

Adli *et al.* found that the wakefield produced by the short proton bunches could accelerate electrons to energies of up to 2 gigaelectronvolts in a plasma that is only about 10 metres in length. For comparison, at the European X-ray free-electron laser facility (European XFEL) in Germany, electrons are accelerated to energies of up to 17.5 gigaelectronvolts in an accelerator that is about 2 km long (see go.nature.com/2n6857t). In addition to providing compact acceleration, the authors' approach has a key advantage over standard accelerators and other wakefield accelerators. Because the proton bunches are stiff, they maintain their structure

and speed. As a result, high-energy electrons can be produced in a single acceleration stage, as opposed to the complex multi-stage process that is needed in other accelerators.

Usually, the higher the energy of a particle beam, the longer it takes to stop (dump) the beam after use. The dumping of high-energy beams has become a serious issue because of the requirement of longer dumping lengths, which in turn increases the production of unwanted radioactive isotopes in the dense materials used for the dumping. The authors show that their accelerated electrons can form a beam of short electron bunches, which would encounter a large collective force if injected into an appropriately prepared plasma. Such a beam could therefore be stopped over a much shorter distance than conventional beams, inducing little radioactivity⁵. Overall, the authors' work represents a major step towards the development of future high-energy particle accelerators that use collective force. ■

Toshiki Tajima is in the Department of Physics and Astronomy, University of California, Irvine, California 92697, USA.
e-mail: tajima@uci.edu

1. Evans, L. & Bryant, P. J. *Instrum* **3**, S08001 (2008).
2. Tajima, T. & Dawson, J. M. *Phys. Rev. Lett.* **43**, 267–270 (1979).
3. Adli, E. *et al.* *Nature* **561**, 363–367 (2018).
4. Gschwendtner, E. *et al.* *Nucl. Instrum. Meth. Phys. Res. A* **829**, 76–82 (2016).
5. Wu, H.-C., Tajima, T., Habs, D., Chao, A. W. & Meyer-ter-Vehn, J. *Phys. Rev. ST Accel. Beams* **13**, 101303 (2010).

CANCER

T cells home in on brain tumours

Immunotherapies activate T cells to destroy tumours, but the approach has failed in some brain cancers. A strategy to improve migration of T cells across the blood–brain barrier could overcome this limitation. [SEE ARTICLE P.331](#)

MICHAEL PLATTEN

Therapies that activate immune cells called T cells to target tumours are an efficient way to combat many types of cancer¹. But an aggressive brain cancer called glioblastoma has proved a particular challenge for immunotherapies². The blood–brain barrier protects the brain against immune-cell infiltration to prevent the potentially life-threatening effects of brain inflammation. This phenomenon is beneficial in normal circumstances, but it prevents T cells from reaching glioblastomas, making the tumours immunologically 'cold'³. On page 331, Samaha and colleagues⁴ report a way to trigger infiltration of T cells into the brains of mice, thus making

glioblastomas vulnerable to immunotherapy.

In the disease encephalitis, brain inflammation occurs because T cells that are typically excluded from the brain migrate across the blood–brain barrier. This migration is a coordinated process that requires activated T cells circulating in the bloodstream to adhere to endothelial cells, which line blood vessels. Adhesion is mediated by the binding of ligand molecules on T cells to cell-adhesion molecules such as ALCAM, ICAM-1 and VCAM-1 on endothelial cells⁵. These cell-adhesion molecules are expressed at higher than normal levels in encephalitis⁶. Binding between ALCAM and the T-cell ligand CD6 halts the progress of activated T cells through blood vessels, allowing subsequent binding by ICAM-1 and VCAM-1.



50 Years Ago

A campaign was opened last week for funds to refloat the Great Britain, one of the three major ships designed by Brunel. The object is to tow her back from the Falkland Islands to the Bristol shipyard ... The Great Britain was the first ocean-going iron ship and the first to be driven by propeller ... Brunel intended the ship to carry passengers of the Great Western Railway ... to New York, but the Great Britain made only a few transatlantic voyages before running aground ... Brunel managed to refloat the ship, which for the next 20 years carried emigrants to Australia ... In 1875, the Great Britain's engines were removed and she was converted to sail, plying between Liverpool and San Francisco until put out of service by a fire near the Falkland Islands ... Despite the ship's age, her structure is still sound enough to survive the journey back to Britain.
From *Nature* 21 September 1968

100 Years Ago

On the afternoon of Saturday, August 24 last, the allotment-holders of a small area in Hendon ... were sheltering in their sheds during a heavy thundershower, when they observed that small fish were being rained to the ground. The fish were precipitated on three adjoining roads and on the allotment-gardens enclosed by the roads; the rain swept them from the roads into the gutters and from the roofs of the sheds ... It is not easy to say how many fish fell, but ... they were numerous ... All the examples which came into my hands ... prove to be the lesser sand-eel (*Ammodytes tobianus*) ... The place where the sand-eels in question were deposited lies about one-quarter of a mile from the seashore ... The only explanation ... is that a shoal of sand-eels was drawn up by a waterspout.
From *Nature* 19 September 1918

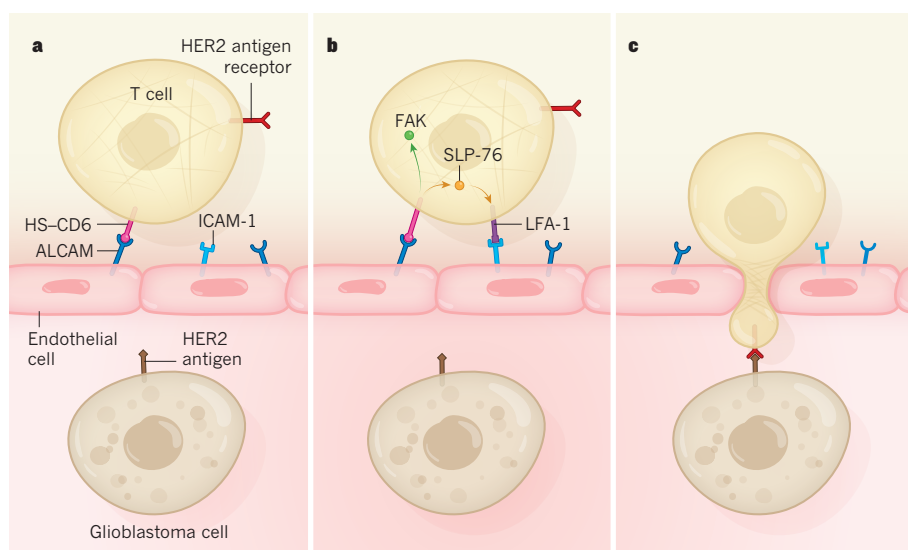


Figure 1 | Targeting of tumour-specific T cells. Immune cells called T cells harbour ligand molecules that bind to the receptor molecules ALCAM and ICAM-1 on endothelial cells, which line blood-vessel walls in the brain. Samaha *et al.*⁴ have developed a strategy to enhance this binding, enabling T cells to cross the blood–brain barrier and infiltrate brain tumours such as glioblastoma. **a**, The authors engineered T cells to express both a synthetic ALCAM-specific ligand, HS-CD6, and an antigen-receptor protein designed to bind to the antigen molecule HER2 on glioblastoma cells. HS-CD6 bound ALCAM with high affinity. **b**, This binding resulted in activation of the protein SLP-76, which induced the ligand LFA-1 to move to the cell surface and bind ICAM-1, strengthening endothelial-cell binding. Binding also led to activation of the protein FAK. **c**, FAK remodelled the actin-protein network that gives the cell its shape (faint lines), enabling the T cell to squeeze between endothelial cells. The HER2 antigen receptor then bound HER2 on glioblastoma cells, triggering an immune response against the tumour.

Once binding of T cells by cell-adhesion molecules reaches a critical threshold, the T cells can migrate between endothelial cells and so out of the vessel into the brain.

In glioblastoma, however, the brain vasculature is reprogrammed such that endothelial cells produce little or no ICAM-1 and VCAM-1 (ref. 7). If researchers could increase adhesion between T cells and endothelial cells in people with glioblastoma, as occurs in encephalitis, it might be possible to enable transendothelial migration of T cells.

Samaha and colleagues found that endothelial cells in glioblastoma overexpress ALCAM. They reasoned that, by engineering T cells to bind to ALCAM more firmly, they could enhance T-cell anchoring in the endothelium and subsequently improve transendothelial migration. To this end, the authors generated a synthetic ligand for ALCAM, derived from CD6. They engineered their molecule, which they named homing-system CD6 (HS-CD6) such that individual ligands interacted with one another to produce a multimeric protein. The researchers introduced the synthetic ligands into T cells using a retrovirus construct. They found that the presence of multimeric HS-CD6 on T cells enhanced adhesiveness between these cells and ALCAM-expressing endothelial cells and, as predicted, enabled transendothelial migration in *in vitro* models.

Samaha and colleagues also uncovered details of the molecular program by which HS-CD6 triggers transendothelial migration.

On binding by ALCAM, HS-CD6 activates the protein SLP-76 in T cells. SLP-76 mobilizes the protein LFA-1, which moves to the cell surface and binds the few ICAM-1 molecules present on endothelial cells, further enhancing binding between T cells and endothelial cells. These changes also activate FAK, a protein that modulates the network of actin proteins that confer T-cell shape, enabling T cells to squeeze between endothelial cells, crossing the blood–brain barrier (Fig. 1).

The next step for the authors was to ensure that T cells entering the brain would home in on tumour cells. T cells harbour antigen-receptor proteins on their surfaces that bind to specific protein fragments called antigens on target cells, enabling T cells to pick out foreign cells for destruction. Samaha *et al.* engineered T cells to express an antigen receptor that was designed to bind to human epidermal growth factor receptor 2 (HER2) — an antigen produced by glioblastoma cells. They then introduced these cells into mice in whose brains human glioblastomas had been surgically implanted. T cells that expressed both HS-CD6 and the HER2-specific antigen receptor infiltrated the glioblastomas, leading to complete remission and long-term survival in most of the treated animals. By contrast, T cells harbouring only the antigen receptor (which are typically used

for cancer immunotherapy) did not infiltrate the tumour.

This study lays out a viable strategy for immunotherapy in glioblastoma. But key challenges must be overcome before we can translate the discovery from mice to patients. For instance, ALCAM is expressed by a variety of cell types, including bone-marrow cells⁸. More studies will be required to assess whether the integrity and functions of these non-endothelial cells are affected by the approach. In addition, toxicity could be an issue if T-cell targeting damages healthy brain tissue, either directly or indirectly. Strategies to limit T-cell activation and lifespan using genetic ‘off’ switches have already been developed⁹, and could potentially be used to combat such toxicity. Encouragingly, the fact that the authors’ mice survived long-term suggests that the treatment did not cause severe toxicity in the animals. However, the group did not investigate the persistence and activity of the HER2-targeting T cells in the body, or examine whether the cells targeted non-glioblastoma cell types.

Finally, targeting T cells to brain tumours is only a first, albeit crucial, step in triggering an effective antitumour T-cell response against glioblastoma. T cells entering a glioblastoma will encounter a profoundly immunosuppressive microenvironment created by low oxygen and pH levels and immunosuppressive molecules. This did not harm T cells in the glioblastoma-harboring mice, but these animals do not mimic many key features of true human glioblastoma — and immunosuppression will certainly pose a challenge in humans. A successful immunotherapy strategy for glioblastoma will ultimately consist of a combinatorial therapy that allows enough active tumour-specific T cells to enter and persist in an immune-permissive tumour microenvironment¹⁰. Such an approach would transform this deadly disease from an immunologically cold target to a hot one. ■

Michael Platten is in the Department of Neurology, Heidelberg University, Medical Faculty Mannheim and the German Cancer Research Center, 69120 Heidelberg, Germany. e-mail: m.platten@dkfz.de

1. Mellman, I., Coukos, G. & Dranoff, G. *Nature* **480**, 480–489 (2011).
2. Platten, M. & Reardon, D. A. *Semin. Neurol.* **38**, 62–72 (2018).
3. Thorsson, V. *et al.* *Immunity* **48**, 812–830 (2018).
4. Samaha, H. *et al.* *Nature* **561**, 331–337 (2018).
5. Engelhardt, B., Vajkoczy, P. & Weller, R. O. *Nature Immunol.* **18**, 123–131 (2017).
6. Ransohoff, R. M. & Engelhardt, B. *Nature Rev. Immunol.* **12**, 623–635 (2012).
7. Quail, D. F. & Joyce, J. A. *Cancer Cell* **13**, 326–341 (2017).
8. Hu, X. *et al.* *Nature Commun.* **7**, 13095 (2016).
9. June, C. H. & Sadelain, M. *N. Engl. J. Med.* **379**, 64–73 (2018).
10. Weller, M. *et al.* *Nature Rev. Neurol.* **13**, 363–374 (2017).

This article was published online on 5 September 2018.

Decoding the phase structure of QCD via particle production at high energy

Anton Andronic^{1,2}, Peter Braun-Munzinger^{1,3,4*}, Krzysztof Redlich^{1,5} & Johanna Stachel³

Recent studies based on lattice Monte Carlo simulations of quantum chromodynamics (QCD)—the theory of strong interactions—have demonstrated that at high temperature there is a phase change from confined hadronic matter to a deconfined quark–gluon plasma in which quarks and gluons can travel distances that greatly exceed the size of hadrons. Here we show that the phase structure of such strongly interacting matter can be decoded by analysing particle production in high-energy nuclear collisions within the framework of statistical hadronization, which accounts for the thermal distribution of particle species. Our results represent a phenomenological determination of the location of the phase boundary of strongly interacting matter, and imply quark–hadron duality at this boundary.

Atomic nuclei are bound by the strong force between their constituent ‘nucleons’: protons and neutrons. Although the density in the centre of a heavy nucleus is extremely large (about 10^{14} times the density of water), the mean distance between nucleons exceeds their diameter (the radius of the nucleon is about $0.88 \text{ fm} = 0.88 \times 10^{-15} \text{ m}$ and the number density inside a nucleus is $n_0 = 0.16 \text{ fm}^{-3}$). Thus, normal nuclear matter is a dilute many-body system. If such matter is compressed or heated in high-energy nuclear collisions (see, for example, refs ^{1–3} to even higher densities or high temperatures (typically of the order of $k_B T \approx m_\pi c^2$, where m_π is the mass of the lightest hadron (the pion), T is the temperature, k_B is Boltzmann’s constant and c is the speed of light), then quarks, the building blocks of nucleons, are expected^{4–7} to be no longer confined but able to move over distances much larger than the size of the nucleon. Such a ‘deconfined’ state of matter, named the quark–gluon plasma (QGP)⁸, is likely to have existed in the early Universe within the first microseconds after the Big Bang⁹. One of the challenging questions in modern nuclear physics is to identify the structure and phases of such strongly interacting matter¹⁰.

Evidence for the existence, in the laboratory, of the QGP has been obtained by studying collisions between heavy atomic nuclei (Au and Pb) at ultra-relativistic energies. The first relevant results came from experiments at the CERN (European Organization for Nuclear Research) Super Proton Synchrotron (SPS) accelerator¹¹. Using the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory (BNL), experiments confirmed the existence of this new state of matter, providing further strong evidence for QGP formation and expansion dynamics in the hot fireball produced in high-energy nuclear collisions. Supporting evidence was also obtained from experiments at the BNL Alternating Gradient Synchrotron (AGS) through the discovery of collective dynamics at high energy¹². For nuclear collisions, the centre-of-mass energies per nucleon pair, $\sqrt{s_{NN}}$, covered by different accelerator facilities are: (1) the BNL AGS, $\sqrt{s_{NN}} = 2.7\text{--}4.8 \text{ GeV}$, (2) the CERN SPS, $\sqrt{s_{NN}} = 6.2\text{--}17.3 \text{ GeV}$, (3) the BNL RHIC, $\sqrt{s_{NN}} = 7.0\text{--}200 \text{ GeV}$ and (4) the CERN Large Hadron Collider (LHC), $\sqrt{s_{NN}} = 2.76\text{--}5.02 \text{ TeV}$.

The results from RHIC showed that the QGP behaves more like a nearly ideal, strongly interacting fluid than a weakly interacting gas of quarks and gluons^{1,3,13–16}. These results were confirmed and extended into hitherto unexplored regions of phase space (in particular, high

transverse momenta) by experiments at the CERN Large Hadron Collider (LHC)^{17–19}. At LHC energies, the fireball formed in Pb–Pb collisions is so hot and dense that quarks or gluons (partons) produced initially with energies of up to a few hundred gigaelectronvolts lose a substantial fraction of their energy while traversing it.

The characterization of the QGP in terms of its equation of state (which expresses pressure as a function of energy density) and of its transport properties (such as its viscosity or diffusion coefficients) as well as delineating the phases of strongly interacting matter²⁰ is a major ongoing research effort^{2,3,18,21,22}. However, it has turned out that direct connections between the underlying theory of strong interactions in the standard model of particle physics, QCD²³ and the experimental data are not readily to be established. This is because the constituents of the QGP—the coloured quarks and gluons—are not observable as free particles, a fundamental property of QCD called ‘confinement’. What is observable are colourless bound states of these partons, resulting in mesons and baryons; these bound states are generally referred to as hadrons. Furthermore, the equations of QCD can be solved analytically only in the high-energy and short-distance limit where perturbative techniques can be used owing to the asymptotic freedom property of QCD^{24,25}. This is unfortunately not possible for the QGP, where typical distance scales exceed the size of the largest atomic nuclei and the typical momentum scale is low. The only technique known at present is lattice QCD (LQCD)²⁶, whereby the QCD equations are solved numerically by discretizing the QCD Lagrangian on a four-dimensional space-time lattice and evaluating them statistically via Monte Carlo methods.

In the following sections we discuss how the phase structure of strongly interacting matter described by LQCD can be decoded by analysing particle production in high-energy nuclear collisions. This is achieved by using the observed thermalization pattern of particle abundances within the framework of statistical hadronization at various collision energies.

Connecting hadronic states and QCD constituents

From LQCD calculations, a deconfinement transition from matter composed of hadronic constituents (that is, hadronic matter) to a QGP has indeed been predicted (see ref. ²⁶ for an early review) at an energy density of about 1 GeV fm^{-3} . Besides deconfinement, there is also a

¹Research Division and EMMI, GSI Helmholtzzentrum für Schwerionenforschung, Darmstadt, Germany. ²Institut für Kernphysik, Universität Münster, Münster, Germany. ³Physikalisches Institut, Universität Heidelberg, Heidelberg, Germany. ⁴Institute of Particle Physics and Key Laboratory of Quark and Lepton Physics (MOE), Central China Normal University, Wuhan, China. ⁵University of Wrocław, Institute of Theoretical Physics, Wrocław, Poland. *e-mail: p.braun-munzinger@gsi.de

chiral symmetry restoration transition expected in high-energy-density matter^{27,28}.

Owing to the very small masses of the up and down quarks, the equations of QCD exhibit symmetries, called chiral symmetries, that allow separate transformations among the right-handed quarks (with spin oriented in the direction of momentum) and left-handed quarks. Such symmetries, however, are not manifest in the observed strongly interacting particles; these particles do not come in opposite-parity pairs. Thus, chiral symmetry must be spontaneously broken at finite energy density. Consequently, QCD predicts the existence of a chiral transition between a phase in which chiral symmetry is broken, at low temperature or density, and a chirally symmetric phase at high temperature or density. The connection between deconfinement and chiral transition is theoretically not fully understood.

It has been demonstrated²⁹, again using the methods of LQCD, that at zero baryo-chemical potential μ_b the deconfinement transition is linked to the restoration of chiral symmetry and that it is of crossover type with a continuous, smooth but rapid increase of thermodynamic quantities in a narrow region around the pseudo-critical temperature T_c . Henceforth using units such that $\hbar = 1$, $k_B = 1$ and $c = 1$, the value of T_c at vanishing μ_b is currently calculated in LQCD to be 154 ± 9 MeV³⁰ and 156 ± 9 MeV^{31,32} with different fermion actions, in excellent agreement (uncertainties quoted here and elsewhere are the standard error of the mean). Recent LQCD results also quantify the small decrease of T_c with increasing μ_b as long as $\mu_b < 3T_c$. Within this parameter range the transition is still of crossover type. A fundamental question is the possible existence of a critical endpoint, where a genuine second-order chiral phase transition is expected. This has been addressed both experimentally (see a review in ref. ³³) and theoretically (see a review in ref. ³⁴) but remains one of the outstanding questions related to our understanding of the phase structure of hot and dense QCD matter.

These results do not shed light on the mechanism of the transition from deconfinement to confinement. In fact, the crossover nature of the chiral transition raises the question whether hadron production from a deconfined medium might also happen over a wide range of temperatures and how confinement can be implemented in a smooth transition without leading to free quarks. A related question is whether colourless bound states (hadrons) might survive in a deconfined medium. The present work attempts to shed light on some of these questions by connecting LQCD phenomenology and the impressive body of results on hadron production in central collisions between two heavy atomic nuclei at high energy. Central collisions are nearly head-on collisions; centrality is calculated in experiments matching measured particle multiplicity or energy to the geometry of the collision (see details in ref. ¹⁷).

The QCD Lagrange density is formulated entirely in terms of the basic constituents of QCD, the quarks and gluons. The masses of hadrons as colourless bound states of quarks and gluons are well calculated within LQCD, showing remarkable agreement with experiment³⁵. This confirms that chiral symmetry is broken in the QCD vacuum, as reflected in the mass differences between parity partners as well as the existence of anomalously light pions as approximate Goldstone bosons associated with spontaneous symmetry breaking.

One of the consequences of confinement in QCD is that physical observables require a representation in terms of hadronic states. Indeed, as has been noted recently in the context of QCD thermodynamics (see ref. ³⁶ and references therein) the corresponding partition function Z can be very well approximated within the framework of the hadron resonance gas, as long as the temperature stays below T_c . To make this more transparent, we first note that all thermodynamic variables, such as pressure P and entropy density can be expressed in terms of derivatives of logarithms of Z . For P , for example, we obtain for a system with volume V and temperature T :

$$\frac{P}{T^4} = \frac{1}{T^3} \frac{\partial \ln[Z(V, T, \mu)]}{\partial V} \quad (1)$$

The results of refs ^{37–42} imply that, as long as $T \leq T_c$,

$$\ln[Z(T, V, \mu)] \approx \sum_{i \in \text{mesons}} \ln[Z_{m_i}(T, V, \mu_Q, \mu_s)] + \sum_{i \in \text{baryons}} \ln[Z_{m_i}(T, V, \mu_b, \mu_Q, \mu_s)] \quad (2)$$

where the partition function of the hadron resonance gas model is expressed in mesonic and baryonic components, where m_i is the mass of a given hadron. The chemical potential μ then reflects the baryonic, electric charge and strangeness components $\mu = (\mu_b, \mu_Q, \mu_s)$.

To make this connection quantitative, detailed investigations have recently been made into the contribution of mesons and baryons to the total pressure of the matter. In particular, in refs ^{36,38} and references therein, the equation of state and different fluctuation observables are evaluated in the hadronic sector via the hadron resonance gas and compared to predictions from LQCD. Very good agreement is obtained for temperatures up to very close to T_c , lending further support to the hadron–parton duality described by equation (2).

The partition function of the hadron resonance gas in equation (2) is evaluated as a mixture of ideal gases of all stable hadrons and resonances. In the spirit of the S-matrix formalism³⁹, which provides a consistent theoretical framework to implement interactions in a dilute many-body system in equilibrium, the presence of resonances corresponds to attractive interactions among hadrons. This is generally a good approximation, because for the temperatures considered here ($T < 165$ MeV) the total particle density is low, $n < 0.5 \text{ fm}^{-3}$.

Sometimes, additional repulsive interactions are modelled with an ‘excluded volume’ prescription (see ref. ³⁷ and references therein), which is inherently a low-density approach. For weak repulsion, implying excluded volume radii of $r_0 < 0.3$ fm, the effect of the correction is mainly to decrease particle densities, whereas the important thermal parameters T and μ_b are little affected. Strong repulsion cannot be modelled that way: much larger r_0 values lead to, among others, unphysical (superluminous) equations of state, in contradistinction to results from LQCD. In the following we use $r_0 = 0.3$ fm for both mesons and baryons. All results on thermal parameters described below are unchanged from what is obtained in the non-interacting limit except for the overall particle density, which is reduced by up to 25%.

Over the course of the past 20 years it has become apparent^{40–45} that the yields of all hadrons produced in central collisions can be very well described by computing particle densities from the hadronic partition function described by equation (2). To obtain particle yields at a particular temperature T_c and μ_b one multiplies the thermal densities obtained in this way with the fireball volume V . In practice, T_c , μ_b and V , the parameters at ‘chemical freeze-out’ from which point on all hadron yields are frozen, are determined from a fit to the experimental data. We note that V is actually the volume that corresponds to a slice of one unit of rapidity y , centred at mid-rapidity. Experimentally, the rapidity density dN/dy of a hadron is obtained by integrating its momentum-space distribution over the momentum component transverse to the beam direction. In general we take these yields N at mid-rapidity ($y = 0$), where the centre of mass of the colliding system is at rest.

As will be discussed below, this ‘statistical hadronization approach’ provides, via equations (1) and (2), a link between data on hadron production in ultra-relativistic nuclear collisions and the QCD partition function. This link may shed light on the QCD phase diagram. The possibility of such a connection was surmised early on^{6,46} and various aspects of it have been discussed more recently^{44,47–53}.

The full power of this link, however, becomes apparent only with the recent precision data from the LHC. Below, we discuss the accuracy that can be achieved in the description of hadron production using the parton–hadron duality concept described by equation (2). We first focus on hadrons that contain only light quarks with flavours up, down and strange (u , d and s) and place emphasis on LHC data. Those show matter and antimatter production in equal amounts, thus indicating that μ_b is very close to zero. It is in this energy region that the LQCD

approach can be applied essentially without approximations using current computer technology. We then explore the lower-energy region (500 GeV $> \sqrt{s_{NN}} > 15$ GeV) and show that consistent information on the QCD phase diagram for $0 < \mu_b < 300$ MeV can be achieved by quantitatively comparing LQCD predictions for finite μ_b with results from statistical hadronization analysis of hadron production data. In section ‘Statistical hadronization of heavy quarks’ we discuss how the statistical hadronization approach can be extended to include heavy (charm c and bottom b) quarks. We further discuss how the recent LHC data can provide information on the hadronization of these heavy quarks during the expansion and cooling of the QGP formed in such high-energy central nuclear collisions.

Statistical hadronization of light quarks

The description of particle production in nucleus–nucleus collisions in the framework of the statistical hadronization approach is particularly transparent at the LHC energy where the chemical freeze-out is quantified, essentially, by the temperature T_{cf} and the volume V of the fireball produced.

The parameters of the statistical hadronization approach are obtained with considerable precision by comparison with the yields of particles measured by the ALICE Collaboration^{54–60}. To match the measurement, the calculations include all contributions from the strong and electromagnetic decays of high-mass resonances. For π^\pm , K^\pm and K^0 mesons, the contributions from heavy-flavour hadron decays are also included. The measurement uncertainty σ is accounted for as the quadratic sum of statistical and systematic uncertainties; see below.

For the most central Pb–Pb collisions, the best description of the ALICE data on yields of particles in one unit of rapidity at mid-rapidity is obtained with $T_{cf} = 156.5 \pm 1.5$ MeV, $\mu_b = 0.7 \pm 3.8$ MeV and $V = 5,280 \pm 410$ fm³. This result is an update of the previous analysis from ref.⁴⁵ using an extended and final dataset. The standard deviations quoted here are exclusively due to experimental uncertainties and do not reflect the systematic uncertainties connected with the model implementation, as discussed below.

Remarkably, the values of the chemical freeze-out temperature $T_{cf} = 156.5 \pm 1.5$ MeV and the pseudo-critical temperature $T_c = 154 \pm 9$ MeV obtained in LQCD agree within errors. This implies that chemical freeze-out takes place close to hadronization of the QGP, lending support also to the hadron–parton duality described by equation (2).

A comparison of the statistical hadronization results obtained with the thermal parameters discussed above and the ALICE data for particle yields is shown in Fig. 1. Impressive overall agreement is obtained between the measured particle yields and the statistical hadronization analysis. The agreement spans nine orders of magnitude in abundance values, encompasses strange and non-strange mesons, baryons including strange and multiply strange hyperons as well as light nuclei and hypernuclei and their antiparticles. A very small value for the baryo-chemical potential $\mu_b = 0.7 \pm 3.8$ MeV, consistent with zero, is obtained, as is expected from the observation of the equal production of matter and antimatter at the LHC⁶¹.

The largest difference between the data and calculations is observed for proton and antiproton yields, where a deviation of 2.7σ is obtained. This difference is connected with an unexpected and puzzling centrality dependence of the ratio $(p + \bar{p})/(\pi^+ + \pi^-)$ (see figure 9 of ref.⁵⁴). As discussed below, the other ratios (hadrons/pions) increase towards more central collisions until a plateau (the grand-canonical limit) is reached. The peculiar behaviour of the $(p + \bar{p})/(\pi^+ + \pi^-)$ ratio at LHC energy is currently not understood. Arguments that this might be connected to annihilation of baryons in the hadronic phase after chemical freeze-out⁶² are not supported by the results of recent measurements of the relative yields of strange baryons to pions⁶³.

A further consequence of the vanishing baryo-chemical potential is that the strangeness chemical potential μ_s also vanishes. This implies that the strangeness quantum number no longer affects the particle production. In the fireball the yield of strange mesons and (multi-) strange baryons is exclusively determined by their mass m and spin degeneracy $(2J + 1)$ in addition to the temperature T .

The thermal origin of all particles including light nuclei and antinuclei is particularly transparent when inspecting how their yields change with particle mass. This is shown in Fig. 2, where the measured yields, normalized to the spin degeneracy, are plotted as a function of the mass m . This demonstrates explicitly that the normalized yields depend exclusively on m and T . For heavy particles ($m \gg T$) without resonance decay contributions their normalized yield simply scales with mass as $m^{3/2} \exp(-m/T)$, illustrated by the nearly linear dependence observed in the logarithmic representation of Fig. 2. We note that, for the subset of light nuclei, the statistical hadronization predictions are not affected by resonance decays. For these nuclei, a small variation in temperature leads to a large variation of the yield, resulting in a relatively precise determination of the freeze-out temperature $T_{nuc} = 159 \pm 5$ MeV, consistent with the value of T_{cf} extracted above.

The incomplete knowledge of the structure and decay probabilities of heavy mesonic and baryonic resonances discussed above leads to systematic uncertainties in the statistical hadronization approach. We note from Fig. 2 that the yields of the measured lightest mesons and baryons (π , K , p and Λ) are substantially increased relative to their primordial thermal production by such decay contributions. For pions, for example, the resonance decay contribution amounts to 70%. For resonance masses larger than 1.5 GeV the individual states start to overlap strongly²³. Consequently, neither their number density nor their decay probabilities can be well determined. Indeed, recent LQCD results indicate that there are missing resonances compared to what is listed in ref.²³. The resulting theoretical uncertainties are difficult to estimate but are expected to be small because T_{cf} is very small compared to the mass of the missing resonances. A conservative estimate is that the resulting systematic uncertainty in T_{cf} is at most 3%. This is consistent with the determination of T_{cf} using only particles whose yields are not influenced by resonance decays (see above). Until now, none of these systematic uncertainties are taken into account in the statistical hadronization analysis described here.

The rapidity densities of light (anti-)nuclei and hypernuclei have been predicted⁶⁴ on the basis of the systematics of hadron production at lower energies. It is nevertheless remarkable that such loosely bound objects (the deuteron binding energy is 2.2 MeV, much less than $T_{nuc} \approx 159$ MeV or $T_{cf} \approx T_c \approx 155$ MeV) are produced with temperatures very close to that of the phase boundary at LHC energy, implying that any further evolution of the fireball has to be close to isentropic. For the hypertriton 3_1H and antihypertriton $^3_1\bar{H}$, the situation is even more dramatic: this object consists of a bound state of (p, n, Λ) , with an energy of only 130 ± 30 keV needed to remove the Λ particle from the bound state. This implies that the Λ particle is very weakly bound to a deuteron, resulting in root-mean-square size for this bound state of close to 10 fm, about the same size as that of the fireball formed in the Pb–Pb collision.

The detailed production mechanism for loosely bound states remains an open question. One, admittedly speculative, possibility is that such objects, at QGP hadronization, are produced as compact, colourless droplets of quark matter with quantum numbers of the final-state hadrons. The concept of possible excitations of nuclear matter into colourless quark droplets has already been considered⁶⁵. In the context of our work, these states should have a lifetime of 5 fm or longer, with excitation energies of 40 MeV or less, for evolution into the final-state hadrons that are measured in the detector. Since by construction they are initially compact, they would also survive a possible short-lived hadronic phase after hadronization. This would be a natural explanation for the striking observation of the thermal pattern for these nuclear bound states emerging from Figs. 1, 2. We note that the observed thermal nature of the production yields of the nuclear bound states is very difficult to reconcile with the assumption that these states are formed by coalescence of baryons, where the yield is proportional to a coalescence factor introduced as the square of the nuclear wavefunction, which varies widely among the various nuclei^{66,67}. For a recent discussion of the application of coalescence models to production of loosely bound states, see ref.⁶⁸.

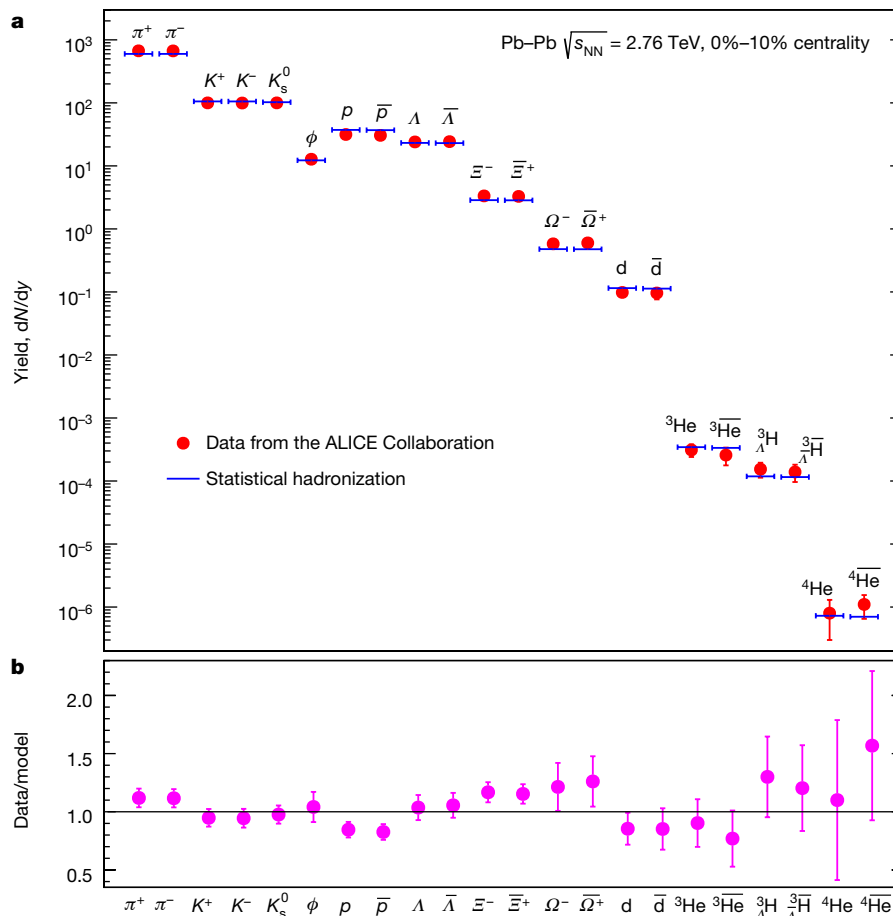


Fig. 1 | Hadron abundances and predictions of the statistical hadronization model. **a**, dN/dy values for different hadrons and nuclei, measured at mid-rapidity (red circles), including the hypertriton ${}^3\text{H}$, are compared with the statistical hadronization analysis (blue bars). The data

are from the ALICE Collaboration for central Pb–Pb collisions at the LHC^{53–59}. **b**, The ratio of the data to statistical hadronization predictions (model), with errors bars determined only from the data as the quadratic sum of statistical and systematic uncertainties.

It could be argued that composite particles such as light nuclei and hypernuclei should not be included in the hadronic partition function described in equation (2). However, all nuclei, including light, loosely bound states, should result from the interaction of the fundamental QCD constituents. This is confirmed by recent LQCD calculations⁶⁹.

The thermal nature of particle production in ultra-relativistic nuclear collisions has been experimentally verified not only at LHC energy, but also at the lower energies of the RHIC, SPS and AGS accelerators. The essential difference is that, at these lower energies, the matter–antimatter symmetry observed at the LHC is lifted, implying non-vanishing values of the chemical potentials. Furthermore, in central collisions at energies below $\sqrt{s_{\text{NN}}} \approx 6$ GeV the cross-section for the production of strange hadrons decreases rapidly, with the result that the average strange hadron yields per collision can be far below unity. In this situation, one needs to implement exact strangeness conservation in the statistical sum in equation (2) and apply the canonical ensemble for the conservation laws^{70,71}. Similar considerations apply for the description of particle yields in peripheral nuclear and elementary collisions. An interesting consequence of exact strangeness conservation is a suppression of strange particle yields when going from central to peripheral nucleus–nucleus collisions or from high multiplicity to low multiplicity events in proton–proton or proton–nucleus collisions. In all cases the suppression is further enhanced with increasing strangeness content of the hadron. Sometimes, additional fugacity parameters g_f are introduced to account for possible non-equilibrium effects of strange- and heavy-flavour hadrons^{44,72}. These parameters modify the thermal yields of particles by factors $g_f^{n_f}$, where the power n_f denotes the number of strange or heavy quarks and antiquarks in the hadron.

Experimental consequences of canonical thermodynamics and strangeness conservation laws have been first seen at SPS energy⁷³. All the above predictions are qualitatively confirmed by the striking results from high-multiplicity proton–proton and p –Pb collisions from the ALICE Collaboration at LHC energy⁶³. The data also explicitly exhibit the plateau in strangeness production for Pb–Pb collisions, which is to be expected when the grand-canonical region is reached, further buttressing the thermal analysis discussed above.

An intriguing observation, first made in ref. ⁷⁴, is that the overall features of hadron production in e^+e^- annihilations resemble that expected from a thermal ensemble with temperature $T \approx 160$ MeV, once exact quantum number conservation is taken into account. In these collisions, quark–antiquark pairs are produced with production yields that are not thermal but are well explained by the electro-weak standard model; see, for example, table 2 in ref. ⁷⁵. Hadrons from these quark pairs (and sometimes gluons) appear as jets in the data. The underlying hadronization process can be well described using statistical hadronization model ideas^{75,76}. These studies reveal further that strangeness production deviates noticeably from a pure thermal production model and that the quantitative description of the measured yields is rather poor. Nevertheless, recognizable thermal features in e^+e^- collisions, where equilibration should be absent, may be a consequence of the generic nature of hadronization in strong interactions.

From a statistical hadronization analysis of all measured hadron yields at various beam energies the detailed energy dependence of the thermal parameters T_{cf} and μ_b has been determined^{41,42,51,77–81}. While μ_b decreases smoothly with increasing energy, the dependence of T_{cf} on energy exhibits a striking feature that is illustrated in Fig. 3: T_{cf} increases with increasing energy (decreasing μ_b) from about 50 MeV

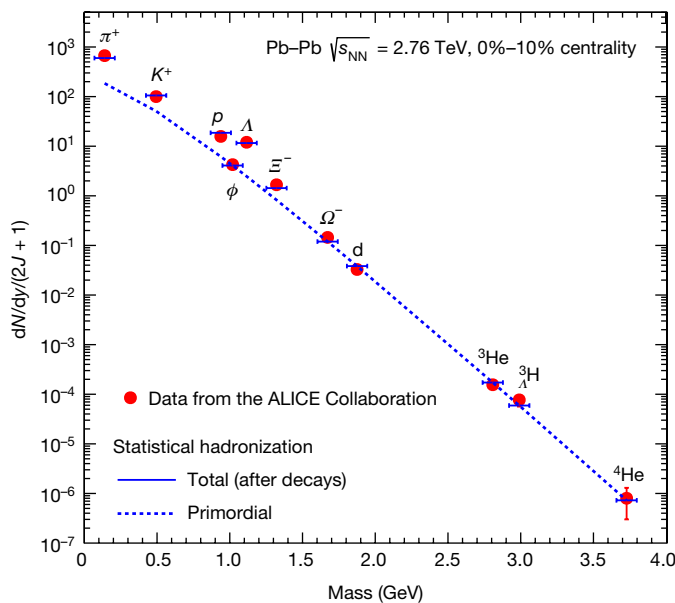


Fig. 2 | Mass dependence of hadron yields compared with predictions of the statistical hadronization model. Only particles (no antiparticles) are included and yields are divided by the spin degeneracy factor $(2J + 1)$. Data are from the ALICE Collaboration for central Pb–Pb collisions at the LHC. For the statistical hadronization approach, the ‘total’ yields (blue bars) include all contributions from high-mass resonances (for the Λ hyperon, the contribution from the electromagnetic decay $\Sigma^0 \rightarrow \Lambda\gamma$, which cannot be resolved experimentally, is also included); the primordial yields before strong and electromagnetic decays are plotted as the dotted line. For more details, see the main text.

to about 160 MeV, where it exhibits a saturation for $\sqrt{s_{NN}} > 20$ GeV. The slight increase of this value compared to $T_{cf} = 156.5$ MeV obtained at LHC energy is due to the inclusion of some data at RHIC energies, but the details of this small difference are currently not fully understood.

The saturation of T_{cf} observed in Fig. 3 lends support to the earlier proposal^{48,50,82} that, at least at high energies, the chemical freeze-out temperature is very close to the QCD hadronization temperature⁵¹, implying a direct connection between data from relativistic nuclear collisions and the QCD phase boundary. This is in accord with the earlier prediction, more than 50 years ago^{83,84}, that hadronic matter cannot be heated beyond this limit. Whether there exists, at lower energies, a critical endpoint⁸⁵ in the QCD phase diagram is currently at the focus of intense theoretical¹⁹ and experimental effort⁷⁷.

To illustrate how well the thermal description of particle production in central nuclear collisions works we show, in Fig. 4, the energy dependence (excitation function) of the relative abundance of several hadron species along with the prediction using the statistical hadronization approach and the smooth evolution of the parameters (see above). Because of the interplay between the energy dependence of T_{cf} and μ_b there are characteristic features in these excitation functions. In particular, maxima appear at slightly different center-of-mass energies in the K^+/π^+ and Λ/π^+ ratios, whereas the corresponding antiparticle ratios exhibit a smooth behaviour⁸⁶.

In the statistical approach in equation (2) and in the Boltzmann approximation, the density $n(\mu_b, T)$ of hadrons carrying baryon number B scales with the chemical potential as $n(\mu_b, T) \propto \exp(B\mu_b/T)$. Consequently, the ratios p/π^+ and d/p , where d refers to a deuteron, scale as $\exp(\mu_b/T)$, whereas the corresponding antiparticle ratios scale as $\exp(-\mu_b/T)$. From Fig. 3, it is apparent that μ_b/T_{cf} decreases with collision energy, accounting for the basic features of particle ratios in the upper panel of Fig. 4. On the other hand, strangeness conservation unambiguously connects, for every T value, the strangeness potential and the baryo-chemical potential, $\mu_s = \mu_b(\mu_b)$. As a consequence, the yields of K^+ and K^- increase and, respectively, decrease with μ_b/T .

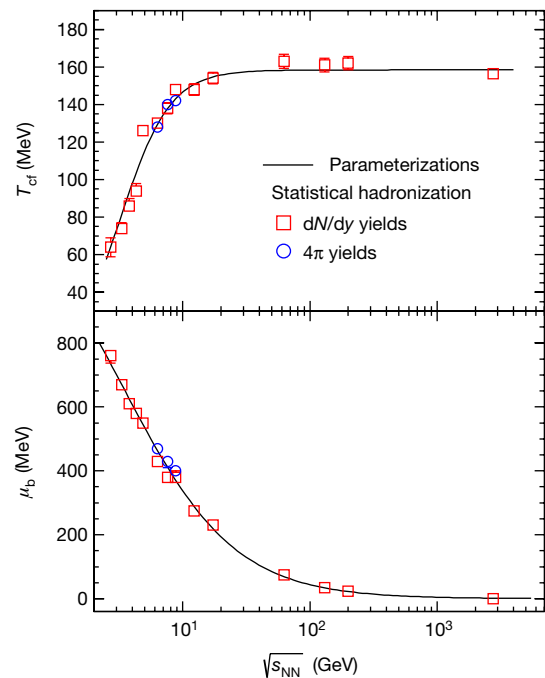


Fig. 3 | Energy dependence of chemical freeze-out parameters T_{cf} and μ_b . The results are obtained from the statistical hadronization analysis of hadron yields (at mid-rapidity, dN/dy , and in full phase space, 4π) for central collisions at different energies. The parameterizations shown are: $T_{cf} = T_{cf}^{lim}/\{1 + \exp[2.60 - \ln(\sqrt{s_{NN}})/0.45]\}$ and $\mu_b = a/(1 + 0.288\sqrt{s_{NN}})$, with $\sqrt{s_{NN}}$ in gigaelectronvolts, $T_{cf}^{lim} = 158.4$ MeV and $a = 1,307.5$ MeV. The uncertainty of the limiting temperature T_{cf}^{lim} , determined from the fit of the five points that represent the highest energies, is 1.4 MeV.

At higher energies, where T and hence pion densities saturate, the Λ/π^+ and K^+/π^+ ratios are decreasing with energy (see lower panel of Fig. 3).

We further note that, for energies beyond that of the LHC, the thermal parameter T_{cf} is determined by the QCD pseudo-critical temperature and the value of μ_b vanishes. Combined with the energy dependence of overall particle production⁸⁷ in central Pb–Pb collisions, this implies that the statistical hadronization model prediction of particle yields at any energy, including those at the possible Future Circular Collider (FCC)⁸⁸ or in ultrahigh-energy cosmic ray collisions⁸⁹, can be made with an estimated precision of better than 15%.

Since the statistical hadronization analysis at each measured energy yields a pair of (T_{cf}, μ_b) values, these points can be used to construct a T versus μ_b diagram, describing phenomenological constraints on the phase boundary between hadronic matter and the QGP; see Fig. 5. We note that the points at low temperature seem to converge towards the value for ground-state nuclear matter ($\mu_b = 931$ MeV). As argued previously⁵², this limit is not necessarily connected to a phase transition. Although the situation at low temperatures and collision energies is complex and at present cannot be investigated with first-principles calculations, the high-temperature, high-collision energy limit allows a quantitative interpretation in terms of fundamental QCD predictions.

The connection between LQCD predictions and experimental chemical freeze-out points is made quantitative in Fig. 5. We use here recent results for the QCD phase boundary from the two leading LQCD groups^{30,90}, represented by the band in Fig. 5. As can be seen, the LQCD values follow the measured μ_b dependence of the chemical freeze-out temperature very closely, demonstrating that with relativistic nuclear collisions one can directly probe the QCD phase boundary between hadronic matter and the QGP. The above results imply that the pseudo-critical temperature of the QCD phase boundary at $\mu_b = 0$ as well as its μ_b dependence up to $\mu_b = 300$ MeV have been determined experimentally. There is indirect but strong evidence from measurements of the initial energy density as well as from hydrodynamical analysis of transverse momentum spectra and from the analysis of jet quenching

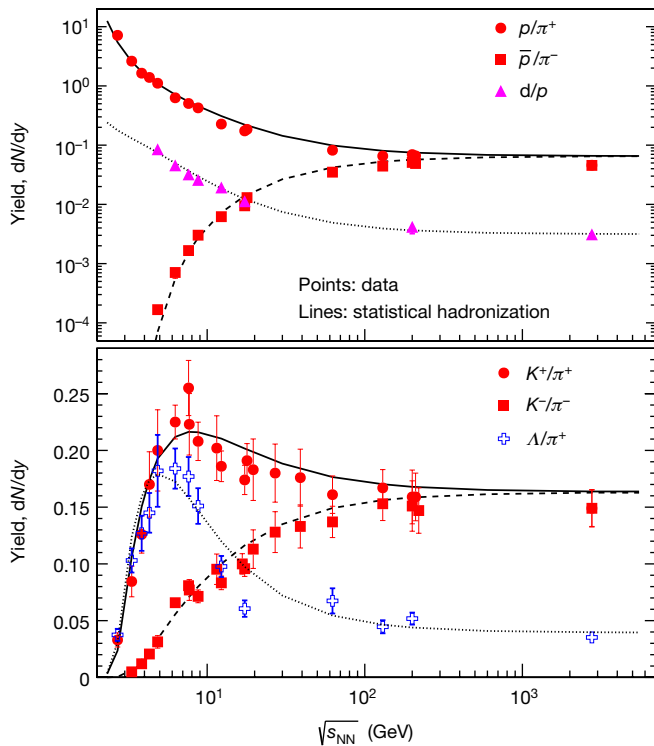


Fig. 4 | Collision-energy dependence of the relative abundance of several hadron species. The data for central collisions (symbols) are compiled from refs ^{77,131}, and are compared to statistical hadronization calculations for the smooth parameterizations of T_{cf} and μ_b as functions of energy shown in Fig. 3.

data that the initial temperatures of the fireball formed in the collision are substantially higher than the values at the phase boundary, reaching 300–600 MeV at RHIC and LHC energies^{91,92}.

The present approach can be extended beyond hadron yields to higher moments of event-by-event particle distributions. Although precision predictions from LQCD exist for higher moments and susceptibilities, especially at LHC energies where μ_b is close to zero (see, for example, refs ^{36,38}), there are formidable difficulties in experimentally determining such higher moments with accuracy. Pioneering experiments were performed at the RHIC accelerator with intriguing but not yet fully conclusive results; for a recent review see ref. ³³. Analyses of higher moments are, therefore, not considered here. Recently, however, the variances (second moments) of strangeness and net-baryon-number fluctuations were reconstructed⁹³ from hadron yields measured in Pb–Pb collisions at mid-rapidity by the CERN ALICE Collaboration. The second moments determined in this way are in impressive agreement with LQCD predictions obtained at the chiral crossover pseudo-critical temperature $T_c \approx 154$ MeV. Furthermore, an interesting strategy was proposed to directly compare experimental data on moments of net-charge fluctuations with LQCD results to identify freeze-out parameters in heavy ion collisions⁹⁴. Within still large systematic uncertainties the extracted freeze-out parameters based on second-order cumulants are consistent with statistical hadronization^{95,96}. Although no formal proof of the above discussed quark–hadron duality near the chiral crossover temperature exists, the empirical evidence has recently clearly been strengthened.

Statistical hadronization of heavy quarks

An interesting question is whether the production of hadrons with heavy quarks can be described with similar statistical hadronization concepts as developed and used in the previous section for light quarks. We note first that the masses of charm and beauty quarks, $m_c = 1.28$ GeV and $m_b = 4.18$ GeV, are much larger than the characteristic scale of QCD, $\Lambda_{QCD} = 332$ MeV for three quark flavours, in the

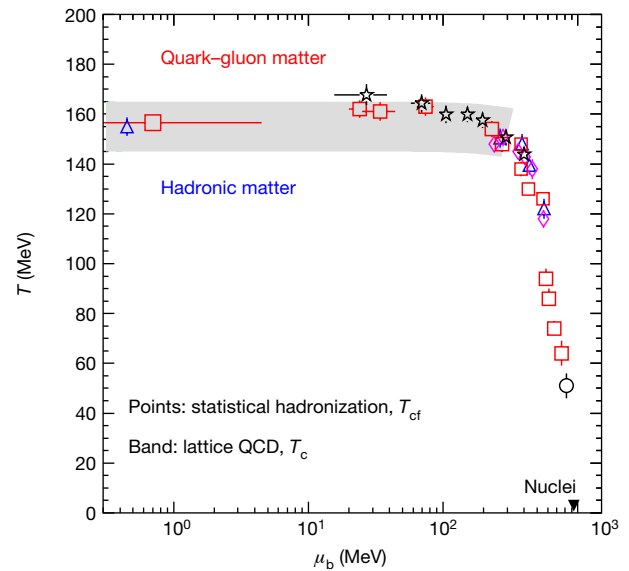


Fig. 5 | Phenomenological phase diagram of strongly interacting matter. The diagram is constructed from chemical freeze-out points that result from statistical hadronization analysis of hadron yields for central collisions at different energies. The freeze-out points, with error bars showing the standard error of the mean, extracted from experimental datasets in our own analysis (red squares) and other similar analyses^{77,78,132,133} (other symbols), are compared to predictions from LQCD^{30,90} (grey band). The inverted triangle marks the value for ground-state nuclear matter (atomic nuclei). The colouring of the labels ‘quark–gluon matter’ and ‘hadronic matter’ indicates the temperature of the matter.

modified minimal subtraction scheme²³. Both masses are also sufficiently larger than the pseudo-critical temperature T_c introduced above, such that thermal production of charm and, in particular, beauty quarks is strongly Boltzmann-suppressed. This is also borne out by quantitative calculations^{97–99}. On the other hand we expect, in particular at LHC energies, copious production of heavy quarks in relativistic nuclear collisions through hard scattering processes which, in view of the large quark masses, can be well described using QCD perturbation theory¹⁰⁰. Consequently, the charm and beauty content of the fireball formed in the nucleus–nucleus collision is determined by the initial hard scattering. Furthermore, the annihilation of charm and beauty quarks is very small, implying that their numbers are tightly preserved during the fireball evolution¹⁰¹.

The charm quarks produced will, therefore, not resemble a chemical equilibrium population for temperature T . However, what is needed for the thermal description proposed below is that the heavy quarks produced in the collision reach a sufficient degree of thermal equilibrium through scattering with the partons of the hot medium. Indeed, the energy loss suffered by energetic heavy quarks in the QGP is indicative of their ‘strong coupling’ with the medium, dominated by light quarks and gluons. The measurements at the LHC^{102,103} and RHIC¹⁰⁴ of the energy loss and hydrodynamic flow of D mesons demonstrate this quantitatively.

Of the various suggested probes of deconfinement, charmonium (the bound state of the charm and anticharm quarks, $c\bar{c}$) is particularly interesting. The J/ψ meson (the ground state of charmonium) is the first hadron for which a clear mechanism of suppression (melting) in the QGP was proposed early on, based on the colour analogue of Debye screening¹⁰⁵. This concept for charmonium suppression was tested experimentally at the SPS accelerator but led, with the turn-on of the RHIC facility, to a number of puzzling results. In particular, the observed rapidity and energy dependence of the suppression ran counter to the theoretical predictions¹⁰⁶.

However, before publication of the first RHIC data, a quarkonium production mechanism based on statistical hadronization, was proposed⁷² that contained a natural explanation for the later observations at RHIC and LHC energy. The basic concept underlying this

statistical hadronization approach⁷² is that charm quarks are produced in initial hard collisions, subsequently thermalize in the QGP and are ‘distributed’ into hadrons at the phase boundary, that is, at chemical freeze-out, with thermal weights as discussed above for the light quarks^{72,101,107,108}. An alternative mechanism for the (re)combination of charm and anticharm quarks into charmonium in a QGP¹⁰⁹ has been proposed, based on kinetic theory. For further developments see refs^{110–113}.

In the statistical hadronization approach, the absence of chemical equilibrium for heavy quarks is accounted for by introducing a fugacity g_c . The parameter g_c is obtained from the balance equation⁷², which accounts for the distribution of all initially produced heavy quarks into hadrons at the phase boundary, with a thermal weight constrained by exact charm conservation. Using this approach, the knowledge of the heavy-quark production cross-section along with the thermal parameters obtained from the analysis of the yields of hadrons composed of light quarks (see section ‘Statistical hadronization of light quarks’) is sufficient to determine the yield of hadrons containing heavy quarks in ultra-relativistic nuclear collisions.

As a consequence, a very useful observable with which to verify the thermal origin of heavy-flavour hadrons produced in a nuclear collision is the abundance ratio of different resonance states such as $\psi(2S)/(J/\psi)$ in the charm-sector or $\Upsilon(2S)/\Upsilon(1S)$ in the bottom-sector (where bottomonium Υ is the bound state of the bottom and antibottom quarks, $b\bar{b}$). Indeed, the first measurement of the $\psi(2S)/(J/\psi)$ abundance ratio at the SPS energy¹¹⁴ demonstrated that there are clearly different production mechanisms for charmonia in elementary and nuclear collisions. This is demonstrated in Fig. 6. Whereas in elementary collisions this ratio is roughly 0.15 and hardly varies with energy, the value observed in central Pb–Pb collisions is more than a factor of four smaller and is remarkably consistent with the assumption that these charmonia are produced at the phase boundary, as are all other hadrons.

The chemical freeze-out temperature barely varies with energy beyond $\sqrt{s_{NN}} = 10$ GeV. Because the charm production cross-section drops out in the $\psi(2S)/(J/\psi)$ ratio, the prediction of the statistical hadronization model for central Pb–Pb collisions at LHC energy is $\psi(2S)/(J/\psi) = 0.035$, with the precision indicated in Fig. 6. Recently, the ALICE Collaboration released the first (transverse momentum integrated) data on the above ratio; the preliminary result is, within the still considerable experimental uncertainties, consistent with the statistical hadronization prediction, lending further support for the thermalization of charm quarks in the hot fireball and the production of charmed hadrons at the phase boundary.

It is also important to assess to what degree the charmonia produced participate in the hydrodynamic expansion of the fireball. This can be done by measuring the second Fourier coefficient of the angular distribution of J/ψ mesons projected onto a plane perpendicular to the beam direction, the so-called elliptic flow. The first measurement of J/ψ elliptic flow at the LHC¹¹⁵ pointed towards rather large values of the elliptic flow coefficients. The recent measurement at $\sqrt{s_{NN}} = 5.02$ TeV from the ALICE Collaboration¹¹⁶ establishes the detailed flow pattern as a function of transverse momentum. The large elliptic flow observed for J/ψ mesons is similar to that observed for open charm mesons¹¹⁷ (mesons that contain one charm quark or anti-quark) and is surprisingly close to the flow coefficients for light hadrons. This provides strong support for charm quark kinetic thermalization, in agreement with the statistical hadronization assumption, and implies that charm quarks couple to the medium in a similar way as do light-flavour quarks. Within the current statistical accuracy the J/ψ data at RHIC are compatible with a null flow signal¹¹⁸. An elliptic anisotropy signal was measured for J/ψ mesons at SPS energy¹¹⁹ and was interpreted as a path-length dependence of the screening.

The response of charmonia produced in ultra-relativistic nuclear collisions to the medium of the fireball is characterized by the nuclear modification factor R_{AA} . This factor is constructed as the ratio between the rapidity densities for J/ψ mesons produced in nucleus–nucleus (AA) collisions and proton–proton collisions, scaled by the number

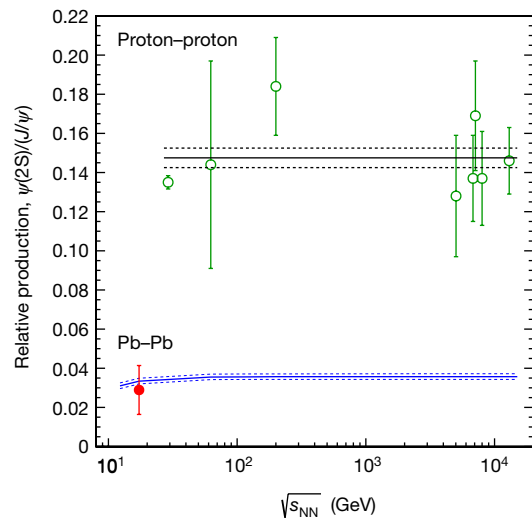


Fig. 6 | Relative production of $\psi(2S)$ and J/ψ mesons as a function of collision energy. The data points for proton–proton collisions are from experiments at SPS, the Hadron–Electron Ring Accelerator (HERA), RHIC and the LHC^{134–139}. The point for central Pb–Pb collisions at the SPS energy is from the NA50 experiment¹¹⁴. The average value of the proton–proton measurements is represented by the black horizontal line with dashed uncertainties. The blue band denotes statistical model calculations for the temperature parameterization from heavy-ion fits; see Fig. 3. The errors on the data are the quadratic sum of statistical and systematic uncertainties.

of nucleon–nucleon collisions in a given centrality bin. Clearly, if all charmonia in the final state originate from hard scattering processes in the initial state, then $R_{AA} = 1$.

In the original colour-screening model¹⁰⁵, R_{AA} was expected to be significantly reduced from unity and to decrease with collision centrality and energy, owing to the increasing energy density of the medium. The early experimental situation until 2009, that is, before LHC turn-on, is summarized in ref.¹²⁰. Indeed, the first data from central Pb–Pb collisions at SPS energy showed a substantial suppression, which could be interpreted in terms of nuclear effects and the Debye screening mechanism. However, the data at RHIC energy exhibited a nearly identical suppression and an unexpected peaking at mid-rapidity¹²⁰, which could not be reconciled with the predictions using the colour-screening model. Both observations on the energy and rapidity dependence of R_{AA} for J/ψ mesons were, however, consistent with their thermal origin^{107,121}, albeit with the qualification of a rather poorly known charm production cross-section.

In the statistical hadronization scenario, the J/ψ nuclear modification factor R_{AA} (see above) is obtained by computing the yields in AA collisions while the yields in proton–proton collisions are taken from experimental data. The R_{AA} value determined in this way should increase with increasing collision energy, implying reduced suppression or even enhancement due to the rapid increase with energy of the charm production cross-section. Clear evidence for such a pattern was obtained with the first ALICE measurements at LHC energy¹²². Since then, a large number of additional data including detailed energy, rapidity, centrality and transverse momentum dependences of R_{AA} for J/ψ as well as hydrodynamic flow and $\psi(2S)/(J/\psi)$ ratio results have provided a firm basis for the statistical hadronization scenario¹⁰¹, with the biggest uncertainties still related to the not-yet-measured value of the open charm cross-section in central Pb–Pb collisions. Current results for J/ψ yields and interpretation within the statistical hadronization picture are summarized in Fig. 7. A large increase of R_{AA} with increasing collision energy is clearly observed. Furthermore, more recent measurements demonstrate (see, for example, figure 16 in ref.¹²³) that the increase is largely concentrated at J/ψ transverse momentum values less than the mass $m_{J/\psi} = 3.1$ GeV. This latter observation was first predicted in ref.¹²⁴. Both provide further support of the original predictions from the statistical hadronization approach.

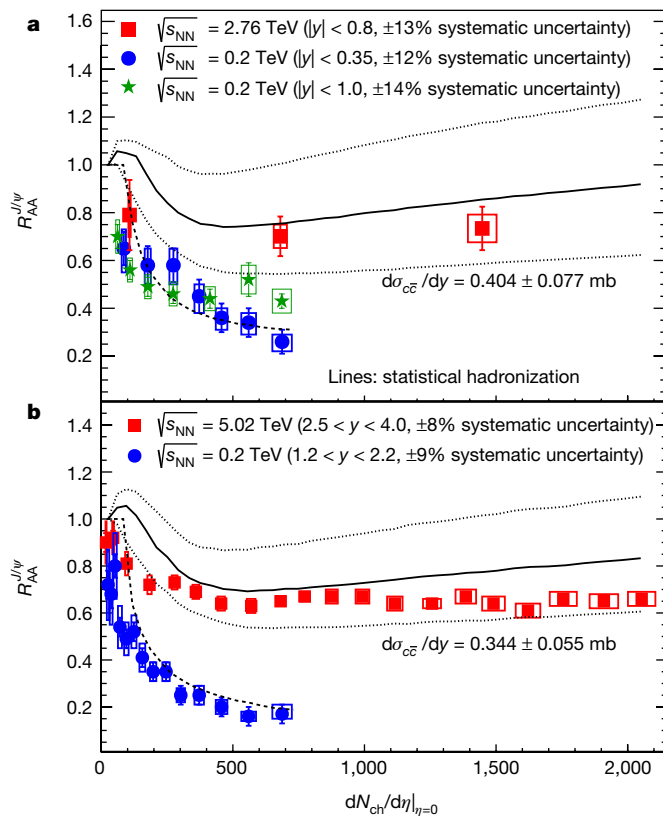


Fig. 7 | The nuclear modification factor R_{AA} for inclusive J/ψ production. a, b, The dependence of R_{AA} on the multiplicity density of charged particles N_{ch} per unit of pseudorapidity η , $dN_{ch}/d\eta|_{\eta=0}$ is shown for mid-rapidity (a) and forward rapidity (b). The data are for Au–Au collisions from the PHENIX Collaboration (blue)^{140,141} and the STAR Collaboration (green)¹⁴² at RHIC and for Pb–Pb collisions from the ALICE Collaboration (red)^{122,123} at the LHC. The open boxes and error bars show the total systematic and statistical uncertainties, respectively.

Recent measurements of the production of bottomonium states at the LHC^{125–127} and at RHIC¹²⁸ can provide further insight into the understanding of the production dynamics of quarkonia in nuclear collisions. The nuclear modification factor for the Υ states exhibits, at LHC energies, a suppression pattern¹²⁶ not unlike that expected in the original Debye screening scenario¹¹². On the other hand, the observed production ratio $\Upsilon(2S)/\Upsilon(1S)$, shown in Fig. 8, is also consistent with a thermalization pattern as one approaches central collisions. Indeed, for central Pb–Pb collisions, this ratio is compatible with the value predicted by the statistical hadronization model for $T \approx 156$ MeV. This provides the tantalizing possibility of adding the bottom flavour as an experimental observable to further constrain the QCD phase boundary with nucleus–nucleus collision data at high energies.

An essential ingredient of the statistical hadronization scenario for heavy quarks is that they can travel, in the QGP, relatively large distances to combine with other uncorrelated partons. The observed increase of the R_{AA} for J/ψ with increasing collision energy strongly supports the notion that the mobility of the heavy quarks is such that it allows travel distances exceeding that of the typical 1-fm hadronic confinement scale. In fact, for LHC energy, the volume of a slice of one unit of rapidity of the fireball exceeds 5,000 fm³, as shown in section ‘Statistical hadronization of light quarks’, which implies that charm quarks can travel distances of the order of 10 fm. This results in the possibility of bound-state formation with all other appropriate partons in the medium having statistical weights quantified by the characteristics of the hadron (mass, quantum numbers) at the phase boundary. The results of the charmonium measurements thereby imply a direct connection to the deconfinement properties of the strongly interacting medium created in ultra-relativistic nuclear collisions.

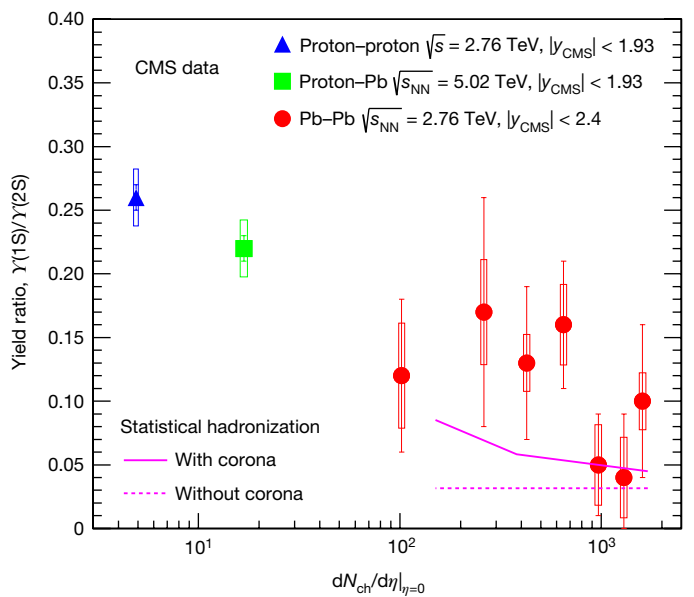


Fig. 8 | Multiplicity dependence of production ratio of bottomonium states $\Upsilon(2S)$ and $\Upsilon(1S)$. The data were measured by the Compact Muon Solenoid (CMS) experiment at the LHC in proton–proton (blue), proton–Pb (green) and Pb–Pb (red) collisions¹²⁵. The magenta lines are statistical hadronization predictions for Pb–Pb collisions; the solid line includes an estimate of the contribution of the production in the corona¹⁰¹ of the colliding nuclei. The open boxes and error bars show the total systematic and statistical uncertainties, respectively.

Outlook

The phenomenological observation of the thermal nature of particle production in heavy-ion collisions at the QCD phase boundary in accord with lattice QCD raises a number of challenging theoretical and experimental issues. An intriguing question is how an isolated quantum system such as a fireball formed in relativistic nuclear collisions can reach an apparently equilibrated state. Similar questions appear^{129,130} in studies of ultracold quantum gases or black holes and may point to a common solution. A second area of interest is the mechanism for the formation of loosely bound nuclear states in a hot fireball at a temperature exceeding their binding energies by orders of magnitude. The question of whether there exist colourless bound states inside a deconfined QGP is related to the experimentally challenging measurements of excited-state populations of quarkonia.

Another priority for the field is the direct observation of the restoration of chiral symmetry and the related critical behaviour in relativistic nuclear collisions with precision measurements and analysis of fluctuation observables. A highlight would be the observation of a critical endpoint in the QCD phase diagram. Making progress with these fundamental issues is at the heart of many ongoing and future theoretical and experimental investigations.

Note added in proof: The proton anomaly discussed in this Review has recently been explained in terms of a pion–nucleon phase-shift analysis¹⁴³. The precision of the cross-over temperature in LQCD has recently been improved to 156.5 ± 1.5 MeV, in very close agreement with the results obtained here¹⁴⁴.

Received: 6 October 2017; Accepted: 3 July 2018;
Published online 19 September 2018.

1. Gyulassy, M. & McLerran, L. New forms of QCD matter discovered at RHIC. *Nucl. Phys. A* **750**, 30–63 (2005).
2. Braun-Munzinger, P. & Stachel, J. The quest for the quark–gluon plasma. *Nature* **448**, 302–309 (2007).
3. Concise review of pre-LHC situation and summary of expectations. Jacak, B. V. & Müller, B. The exploration of hot nuclear matter. *Science* **337**, 310–314 (2012).
4. Itoh, N. Hydrostatic equilibrium of hypothetical quark stars. *Prog. Theor. Phys.* **44**, 291–292 (1970).

5. Collins, J. C. & Perry, M. Superdense matter: neutrons or asymptotically free quarks? *Phys. Rev. Lett.* **34**, 1353–1356 (1975).
6. Cabibbo, N. & Parisi, G. Exponential hadronic spectrum and quark liberation. *Phys. Lett. B* **59**, 67–69 (1975).
7. Chapline, G. & Nauenberg, M. Asymptotic freedom and the baryon-quark phase transition. *Phys. Rev. D* **16**, 450–456 (1977).
8. Shuryak, E. V. Quark-gluon plasma and hadronic production of leptons, photons and pions. *Phys. Lett. B* **78**, 150–153 (1978).
9. Boyanovsky, D., de Vega, H. & Schwarz, D. Phase transitions in the early and the present universe. *Annu. Rev. Nucl. Part. Sci.* **56**, 441–500 (2006).
10. Rajagopal, K. & Wilczek, F. in *At The Frontier Of Particle Physics. Handbook of QCD* Vol. 3 (ed. Shifman, M.) 2061–2151 (World Scientific, New Jersey, 2001).
11. Heinz, U. W. & Jacob, M. Evidence for a new state of matter: an assessment of the results from the CERN lead beam program. Preprint at <https://arxiv.org/abs/nucl-th/0002042> (2000).
12. E877 Collaboration. Observation of anisotropic event shapes and transverse flow in Au + Au collisions at AGS energy. *Phys. Rev. Lett.* **73**, 2532–2535 (1994).
13. STAR Collaboration. Experimental and theoretical challenges in the search for the quark gluon plasma: the STAR Collaboration's critical assessment of the evidence from RHIC collisions. *Nucl. Phys. A* **757**, 102–183 (2005).
14. BRAHMS Collaboration. Quark gluon plasma and color glass condensate at RHIC? The perspective from the BRAHMS experiment. *Nucl. Phys. A* **757**, 1–27 (2005).
15. PHENIX Collaboration. Formation of dense partonic matter in relativistic nucleus-nucleus collisions at RHIC: experimental evaluation by the PHENIX Collaboration. *Nucl. Phys. A* **757**, 184–283 (2005).
16. PHOBOS Collaboration. The PHOBOS perspective on discoveries at RHIC. *Nucl. Phys. A* **757**, 28–101 (2005).
17. Müller, B., Schukraft, J. & Wyslouch, B. First results from Pb+Pb collisions at the LHC. *Annu. Rev. Nucl. Part. Sci.* **62**, 361–386 (2012).
18. Schukraft, J. Heavy ion physics at the Large Hadron Collider: what is new? What is next? *Phys. Scr. T* **158**, 014003 (2013).
19. Braun-Munzinger, P., Koch, V., Schäfer, T. & Stachel, J. Properties of hot and dense matter from relativistic heavy ion collisions. *Phys. Rep.* **621**, 76–126 (2016).
20. Braun-Munzinger, P. & Wambach, J. The phase diagram of strongly-interacting matter. *Rev. Mod. Phys.* **81**, 1031–1050 (2009).
21. Müller, B. Investigation of hot QCD matter: theoretical aspects. *Phys. Scr. T* **158**, 014004 (2013).
22. Satz, H. Probing the states of matter in QCD. *Int. J. Mod. Phys. A* **28**, 1330043 (2013).
23. Particle Data Group Collaboration. Review of particle physics. *Chin. Phys. C* **40**, 100001 (2016).
24. Gross, D. J. & Wilczek, F. Ultraviolet behavior of nonabelian gauge theories. *Phys. Rev. Lett.* **30**, 1343–1346 (1973).
25. Politzer, H. D. Reliable perturbative results for strong interactions? *Phys. Rev. Lett.* **30**, 1346–1349 (1973).
26. Karsch, F. Lattice QCD at high temperature and density. *Lect. Notes Phys.* **583**, 209–249 (2002).
27. Wilczek, F. QCD made simple. *Phys. Today* **53**, 22–28 (2000).
28. Bazavov, A. et al. The chiral and deconfinement aspects of the QCD transition. *Phys. Rev. D* **85**, 054503 (2012).
29. Aoki, Y., Endrodi, G., Fodor, Z., Katz, S. & Szabo, K. The order of the quantum chromodynamics transition predicted by the standard model of particle physics. *Nature* **443**, 675–678 (2006).
30. HotQCD Collaboration. The equation of state in (2+1)-flavor QCD. *Phys. Rev. D* **90**, 094503 (2014).
31. Borsányi, S. et al. Is there still any T_c mystery in lattice QCD? Results with physical masses in the continuum limit III. *J. High Energy Phys.* **9**, 73 (2010).
32. Borsányi, S. et al. Full result for the QCD equation of state with 2+1 flavors. *Phys. Lett. B* **730**, 99–104 (2014).
33. Luo, X. & Xu, N. Search for the QCD critical point with fluctuations of conserved quantities in relativistic heavy-ion collisions at RHIC: an overview. *Nucl. Sci. Tech.* **28**, 112 (2017).
34. Karsch, F. The last word(s) on CPOD 2013. *Proc. Sci.* **185**, 46 (2013).
35. Dürr, S. et al. Ab-initio determination of light hadron masses. *Science* **322**, 1224–1227 (2008).
36. Bazavov, A. et al. The QCD equation of state to $\mathcal{O}(\mu_B^6)$ from lattice QCD. *Phys. Rev. D* **95**, 054504 (2017).
37. Andronic, A., Braun-Munzinger, P., Stachel, J. & Winn, M. Interacting hadron resonance gas meets lattice QCD. *Phys. Lett. B* **718**, 80–85 (2012).
38. Karsch, F. Thermodynamics of strong interaction matter from lattice QCD and the hadron resonance gas model. *Acta Phys. Polon. B* **7**, 117–126 (2014).
39. Dashen, R., Ma, S.-K. & Bernstein, H. J. S Matrix formulation of statistical mechanics. *Phys. Rev.* **187**, 345–370 (1969).
40. Cleymans, J. & Satz, H. Thermal hadron production in high-energy heavy ion collisions. *Z. Phys. C* **57**, 135–147 (1993).
41. Braun-Munzinger, P., Stachel, J., Wessels, J. & Xu, N. Thermal equilibration and expansion in nucleus-nucleus collisions at the AGS. *Phys. Lett. B* **344**, 43–48 (1995).
42. Braun-Munzinger, P., Redlich, K. & Stachel, J. in *Quark Gluon Plasma* Vol. 3 (eds Hwa, R. C. & Wang, X.-N.) 491–599 (World Scientific, Singapore, 2004).
43. Braun-Munzinger, P., Magestro, D., Redlich, K. & Stachel, J. Hadron production in Au–Au collisions at RHIC. *Phys. Lett. B* **518**, 41–46 (2001).
44. Letessier, J. & Rafelski, J. Hadron production and phase changes in relativistic heavy ion collisions. *Eur. Phys. J. A* **35**, 221–242 (2008).
45. Stachel, J., Andronic, A., Braun-Munzinger, P. & Redlich, K. Confronting LHC data with the statistical hadronization model. *J. Phys. Conf. Ser.* **509**, 012019 (2014).
46. Hagedorn, R. How we got to QCD matter from the hadron side by trial and error. *Lect. Notes Phys.* **221**, 53–76 (1985).
47. Cleymans, J. & Redlich, K. Unified description of freezeout parameters in relativistic heavy ion collisions. *Phys. Rev. Lett.* **81**, 5284–5286 (1998).
48. Stock, R. The parton to hadron phase transition observed in Pb+Pb collisions at 158-GeV per nucleon. *Phys. Lett. B* **456**, 277–282 (1999).
49. Braun-Munzinger, P. & Stachel, J. Particle ratios, equilibration, and the QCD phase boundary. *J. Phys. G* **28**, 1971–1976 (2002).
50. Braun-Munzinger, P., Stachel, J. & Wetterich, C. Chemical freezeout and the QCD phase transition temperature. *Phys. Lett. B* **596**, 61–69 (2004).
51. Andronic, A., Braun-Munzinger, P. & Stachel, J. Thermal hadron production in relativistic nuclear collisions: the hadron mass spectrum, the horn, and the QCD phase transition. *Phys. Lett. B* **673**, 142–145 (2009).
52. Floorchinger, S. & Wetterich, C. Chemical freeze-out in heavy ion collisions at large baryon densities. *Nucl. Phys. A* **890–891**, 11–24 (2012).
53. Bazavov, A. et al. Freeze-out conditions in heavy ion collisions from QCD thermodynamics. *Phys. Rev. Lett.* **109**, 192302 (2012).
54. ALICE Collaboration. Centrality dependence of π , K , p production in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Rev. C* **88**, 044910 (2013).
55. ALICE Collaboration. K_S^0 and Λ production in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Rev. Lett.* **111**, 222301 (2013).
56. ALICE Collaboration. Multi-strange baryon production at mid-rapidity in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Lett. B* **728**, 216–227 (2014).
57. ALICE Collaboration. $K^*(892)^0$ and $\phi(1020)$ production in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Rev. C* **91**, 024609 (2015).
58. ALICE Collaboration. ^3H and ^3He production in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Lett. B* **754**, 360–372 (2016).
59. ALICE Collaboration. Production of light nuclei and anti-nuclei in pp and Pb-Pb collisions at energies available at the CERN Large Hadron Collider. *Phys. Rev. C* **93**, 024917 (2016).
60. ALICE Collaboration. Production of ^4He and ^4He in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV at the LHC. *Nucl. Phys. A* **971**, 1–20 (2018).
61. ALICE Collaboration. Pion, kaon, and proton production in central Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Rev. Lett.* **109**, 252301 (2012).
62. Becattini, F., Grossi, E., Bleicher, M., Steinheimer, J. & Stock, R. Centrality dependence of hadronization and chemical freeze-out conditions in heavy ion collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Rev. C* **90**, 054907 (2014).
63. ALICE Collaboration. Enhanced production of multi-strange hadrons in high-multiplicity proton-proton collisions. *Nat. Phys.* **13**, 535–539 (2017).
64. Andronic, A., Braun-Munzinger, P., Stachel, J. & Stöcker, H. Production of light nuclei, hypernuclei and their antiparticles in relativistic nuclear collisions. *Phys. Lett. B* **697**, 203–207 (2011).
65. Chapline, G. & Kerman, A. *On the Possibility of Making Quark Matter in Nuclear Collisions*. Report No. MIT-CTP-695, <https://inspirehep.net/record/134446/files/CTP-695.pdf> (MIT Center of Theoretical Physics, 1978).
66. Csernai, L. P. & Kapusta, J. I. Entropy and cluster production in nuclear collisions. *Phys. Rep.* **131**, 223–318 (1986).
67. Hirenzaki, S., Suzuki, T. & Tanihata, I. A general formula of the coalescence model. *Phys. Rev. C* **48**, 2403–2408 (1993).
68. ExHIC Collaboration. Exotic hadrons from heavy ion collisions. *Prog. Part. Nucl. Phys.* **95**, 279–322 (2017).
69. NPLQCD Collaboration. Light nuclei and hypernuclei from quantum chromodynamics in the limit of SU(3) flavor symmetry. *Phys. Rev. D* **87**, 034506 (2013).
70. Hagedorn, R. & Redlich, K. Statistical thermodynamics in relativistic particle and ion physics: canonical or grand canonical? *Z. Phys. C* **27**, 541–551 (1985).
71. Hamieh, S., Redlich, K. & Tounsi, A. Canonical description of strangeness enhancement from p-A to Pb-Pb collisions. *Phys. Lett. B* **486**, 61–66 (2000).
72. Braun-Munzinger, P. & Stachel, J. (Non)thermal aspects of charmonium production and a new look at J/ψ suppression. *Phys. Lett. B* **490**, 196–202 (2000).
73. NA57 Collaboration. Energy dependence of hyperon production in nucleus nucleus collisions at SPS. *Phys. Lett. B* **595**, 68–74 (2004).
74. Becattini, F. A thermodynamical approach to hadron production in e^+e^- collisions. *Z. Phys. C* **69**, 485–492 (1996).
75. Becattini, F., Castorina, P., Manninen, J. & Satz, H. The thermal production of strange and non-strange hadrons in e^+e^- collisions. *Eur. Phys. J. C* **56**, 493–510 (2008).
76. Andronic, A., Beutler, F., Braun-Munzinger, P., Redlich, K. & Stachel, J. Thermal description of hadron production in e^+e^- collisions revisited. *Phys. Lett. B* **675**, 312–318 (2009).
77. STAR Collaboration. Bulk properties of the medium produced in relativistic heavy-ion collisions from the beam energy scan program. *Phys. Rev. C* **96**, 044904 (2017).
78. Cleymans, J., Oeschler, H. & Redlich, K. Influence of impact parameter on thermal description of relativistic heavy ion collisions at (1–2)A GeV. *Phys. Rev. C* **59**, 1663–1673 (1999).

Establishing the statistical hadronization model in the RHIC era.

79. Braun-Munzinger, P., Heppe, I. & Stachel, J. Chemical equilibration in Pb + Pb collisions at the SPS. *Phys. Lett. B* **465**, 15–20 (1999).
80. Manninen, J. & Becattini, F. Chemical freeze-out in ultra-relativistic heavy ion collisions at $\sqrt{s_{NN}} = 130$ and 200 GeV. *Phys. Rev. C* **78**, 054901 (2008).
81. STAR Collaboration. Identified particle production, azimuthal anisotropy, and interferometry measurements in Au+Au collisions at $\sqrt{s_{NN}} = 9.2$ GeV. *Phys. Rev. C* **81**, 024911 (2010).
82. Braun-Munzinger, P. & Stachel, J. Dynamics of ultrarelativistic nuclear collisions with heavy beams: an experimental overview. *Nucl. Phys. A* **638**, 3c–18c (1998).
- Establishing the boundary line for chemical freeze-out.**
83. Hagedorn, R. Statistical thermodynamics of strong interactions at high-energies. *Nuovo Cim.* **3** (Suppl.), 147–186 (1965).
84. Hagedorn, R. *Miscellaneous Elementary Remarks about the Phase Transition from a Hadron Gas to a Quark-Gluon Plasma*. Report No. CERN-TH.4100, <http://cds.cern.ch/record/158166/files/198504017.pdf> (CERN, 1985).
85. Stephanov, M. A., Rajagopal, K. & Shuryak, E. V. Signatures of the tricritical point in QCD. *Phys. Rev. Lett.* **81**, 4816–4819 (1998).
86. Braun-Munzinger, P., Cleymans, J., Oeschler, H. & Redlich, K. Maximum relative strangeness content in heavy ion collisions around 30 GeV/A. *Nucl. Phys. A* **697**, 902–912 (2002).
87. ALICE Collaboration. Centrality dependence of the charged-particle multiplicity density at midrapidity in Pb-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV. *Phys. Rev. Lett.* **116**, 222302 (2016).
88. Dainese, A. et al. Heavy ions at the Future Circular Collider. *CERN Yellow Rep.* **3**, 635–691 (2017); <https://e-publishing.cern.ch/index.php/CYRM/article/view/515/371>.
89. Pierre Auger Collaboration. Ultra-high energy cosmic rays: recent results and future plans of Auger. *AIP Conf. Proc.* **1852**, 040001 (2017).
90. Borsanyi, S. et al. QCD equation of state at nonzero chemical potential: continuum results with physical quark masses at order μ^2 . *J. High Energy Phys.* **8**, 53 (2012).
91. PHENIX Collaboration. Enhanced production of direct photons in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV and implications for the initial temperature. *Phys. Rev. Lett.* **104**, 132301 (2010).
92. ALICE Collaboration. Direct photon production in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Lett. B* **754**, 235–248 (2016).
93. Braun-Munzinger, P., Kalweit, A., Redlich, K. & Stachel, J. Confronting fluctuations of conserved charges in central nuclear collisions at the LHC with predictions from lattice QCD. *Phys. Lett. B* **747**, 292–298 (2015).
94. Karsch, F. Determination of freeze-out conditions from lattice QCD calculations. *Cent. Eur. J. Phys.* **10**, 1234–1237 (2012).
95. Borsanyi, S. et al. Freeze-out parameters from electric charge and baryon number fluctuations: is there consistency? *Phys. Rev. Lett.* **113**, 052301 (2014).
96. PHENIX Collaboration. Measurement of higher cumulants of net-charge multiplicity distributions in Au+Au collisions at $\sqrt{s_{NN}} = 7.7$ –200 GeV. *Phys. Rev. C* **93**, 011901 (2016).
97. Braun-Munzinger, P. & Redlich, K. Charmonium production from the secondary collisions at LHC energy. *Eur. Phys. J. C* **16**, 519–525 (2000).
98. Zhang, B.-W., Ko, C.-M. & Liu, W. Thermal charm production in a quark-gluon plasma in Pb-Pb collisions at $\sqrt{s_{NN}} = 25.5$ TeV. *Phys. Rev. C* **77**, 024901 (2008).
99. Zhou, K., Chen, Z., Greiner, C. & Zhuang, P. Thermal charm and charmonium production in quark gluon plasma. *Phys. Lett. B* **758**, 434–439 (2016).
100. Cacciari, M. et al. Theoretical predictions for charm and bottom production at the LHC. *J. High Energy Phys.* **10**, 137 (2012).
101. Andronic, A., Braun-Munzinger, P., Redlich, K. & Stachel, J. Statistical hadronization of heavy quarks in ultra-relativistic nucleus-nucleus collisions. *Nucl. Phys. A* **789**, 334–356 (2007). **Working out predictions for charmonium and bottomonium production at collider energies.**
102. ALICE Collaboration. Suppression of high transverse momentum D mesons in central Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *J. High Energy Phys.* **9**, 112 (2012).
103. ALICE Collaboration. D meson elliptic flow in noncentral Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Rev. Lett.* **111**, 102301 (2013).
104. STAR Collaboration. Observation of D^0 meson nuclear modifications in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV. *Phys. Rev. Lett.* **113**, 142301 (2014).
105. Matsui, T. & Satz, H. J/ψ suppression by quark-gluon plasma formation. *Phys. Lett. B* **178**, 416–422 (1986).
- Debye screening of J/ψ mesons in a QGP.**
106. Vogt, R. J/ψ production and suppression. *Phys. Rep.* **310**, 197–260 (1999).
107. Braun-Munzinger, P. & Stachel, J. in *Relativistic Heavy Ion Physics* (ed. Stock, R.) 424–444 (Landolt-Börnstein – Group I: Elementary Particles, Nuclei and Atoms Vol. 23, Springer, Berlin, 2010).
108. Andronic, A., Braun-Munzinger, P., Redlich, K. & Stachel, J. The thermal model on the verge of the ultimate test: particle production in Pb-Pb collisions at the LHC. *J. Phys. G* **38**, 124081 (2011).
109. Thews, R. L., Schroedter, M. & Rafelski, J. Enhanced J/ψ production in deconfined quark matter. *Phys. Rev. C* **63**, 054905 (2001).
- Continuous formation and destruction of charmonia in the QGP.**
110. Liu, Y.-P., Qu, Z., Xu, N. & Zhuang, P.-F. J/ψ transverse momentum distribution in high energy nuclear collisions at RHIC. *Phys. Lett. B* **678**, 72–76 (2009).
111. Grandchamp, L., Rapp, R. & Brown, G. E. In medium effects on charmonium production in heavy ion collisions. *Phys. Rev. Lett.* **92**, 212301 (2004).
112. Emerick, A., Zhao, X. & Rapp, R. Bottomonia in the quark-gluon plasma and their production at RHIC and LHC. *Eur. Phys. J. A* **48**, 72 (2012).
113. Zhou, K., Xu, N., Xu, Z. & Zhuang, P. Medium effects on charmonium production at ultrarelativistic energies available at the CERN Large Hadron Collider. *Phys. Rev. C* **89**, 054911 (2014).
114. NA50 Collaboration. ψ' production in Pb-Pb collisions at 158 GeV/nucleon. *Eur. Phys. J. C* **49**, 559–567 (2007).
115. ALICE Collaboration. J/ψ elliptic flow in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Rev. Lett.* **111**, 162301 (2013).
116. ALICE Collaboration. J/ψ elliptic flow in Pb-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV. *Phys. Rev. Lett.* **119**, 242301 (2017).
117. ALICE Collaboration. D-meson azimuthal anisotropy in mid-central Pb-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV. *Phys. Rev. Lett.* **120**, 102301 (2018).
118. STAR Collaboration. Measurement of J/ψ azimuthal anisotropy in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV. *Phys. Rev. Lett.* **111**, 052301 (2013).
119. NA50 Collaboration. J/ψ azimuthal anisotropy relative to the reaction plane in Pb-Pb collisions at 158 GeV per nucleon. *Eur. Phys. J. C* **61**, 853–858 (2009).
120. Kluber, L. & Satz, H. in *Relativistic Heavy Ion Physics* (ed. Stock, R.) 373–423 (Landolt-Börnstein – Group I: Elementary Particles, Nuclei and Atoms Vol. 23, Springer, Berlin, 2010).
121. Andronic, A., Braun-Munzinger, P., Redlich, K. & Stachel, J. Evidence for charmonium generation at the phase boundary in ultra-relativistic nuclear collisions. *Phys. Lett. B* **652**, 259–261 (2007).
122. ALICE Collaboration. Centrality, rapidity and transverse momentum dependence of J/ψ suppression in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Lett. B* **743**, 314–327 (2014).
- First experimental evidence of reduced J/ψ suppression at LHC energy.**
123. ALICE Collaboration. J/ψ suppression at forward rapidity in Pb-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV. *Phys. Lett. B* **766**, 212–224 (2017).
124. Zhao, X. & Rapp, R. Medium modifications and production of charmonia at LHC. *Nucl. Phys. A* **859**, 114–125 (2011).
125. CMS Collaboration. Event activity dependence of $\Upsilon(nS)$ production in $\sqrt{s_{NN}} = 5.02$ TeV pPb and $\sqrt{s_{NN}} = 2.76$ TeV pp collisions. *J. High Energy Phys.* **4**, 103 (2014).
126. CMS Collaboration. Observation of sequential Υ suppression in PbPb collisions. *Phys. Rev. Lett.* **109**, 222301 (2012).
127. ALICE Collaboration. Suppression of $\Upsilon(1S)$ at forward rapidity in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. *Phys. Lett. B* **738**, 361–372 (2014).
128. PHENIX Collaboration. Measurement of $\Upsilon(1S + 2S + 3S)$ production in p+p and Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV. *Phys. Rev. C* **91**, 024913 (2015).
129. Rigol, M., Dunjko, V. & Olshanii, M. Thermalization and its mechanism for generic isolated quantum systems. *Nature* **452**, 854–858 (2008).
130. Gring, M. et al. Relaxation and prethermalization in an isolated quantum system. *Science* **337**, 1318–1322 (2012).
131. Andronic, A. An overview of the experimental study of quark-gluon matter in high-energy nucleus-nucleus collisions. *Int. J. Mod. Phys. A* **29**, 1430047 (2014).
132. Vovchenko, V., Begun, V. V. & Gorenstein, M. I. Hadron multiplicities and chemical freeze-out conditions in proton-proton and nucleus-nucleus collisions. *Phys. Rev. C* **93**, 064906 (2016).
133. Becattini, F., Steinheimer, J., Stock, R. & Bleicher, M. Hadronization conditions in relativistic nuclear collisions and the QCD pseudo-critical line. *Phys. Lett. B* **764**, 241–246 (2017).
134. NA51 Collaboration. J/ψ , ψ' and Drell-Yan production in pp and pd interactions at 450 GeV/c. *Phys. Lett. B* **438**, 35–40 (1998).
135. HERA-B Collaboration. A measurement of the ψ' to J/ψ production ratio in 920-GeV proton-nucleus interactions. *Eur. Phys. J. C* **49**, 545–558 (2007).
136. PHENIX Collaboration. Measurement of the relative yields of $\psi(2S)$ to $\psi(1S)$ mesons produced at forward and backward rapidity in p+p, p+Al, p+Au and 3He+Au collisions at $\sqrt{s_{NN}} = 200$ GeV. *Phys. Rev. C* **95**, 034904 (2017).
137. LHCb Collaboration. Measurement of J/ψ production in pp collisions at $\sqrt{s} = 7$ TeV. *Eur. Phys. J. C* **71**, 1645 (2011).
138. LHCb Collaboration. Measurement of $\psi(2S)$ meson production in pp collisions at $\sqrt{s} = 7$ TeV. *Eur. Phys. J. C* **72**, 2100 (2012).
139. ALICE Collaboration. Energy dependence of forward-rapidity J/ψ and $\psi(2S)$ production in pp collisions at the LHC. *Eur. Phys. J. C* **77**, 392 (2017).
140. PHENIX Collaboration. J/ψ production vs centrality, transverse momentum, and rapidity in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV. *Phys. Rev. Lett.* **98**, 232301 (2007).
141. PHENIX Collaboration. J/ψ suppression at forward rapidity in Au+Au collisions at $\sqrt{s_{NN}} = 200$ GeV. *Phys. Rev. C* **84**, 054912 (2011).
142. STAR Collaboration. J/ψ production at low p_T in Au+Au and Cu+Cu collisions at $\sqrt{s_{NN}} = 200$ GeV at STAR. *Phys. Rev. C* **90**, 024906 (2014).
143. Andronic, A. et al. The thermal proton yield anomaly in Pb-Pb collisions at the LHC and its resolution. Preprint at <https://arxiv.org/abs/1808.03102> (2018).
144. Steinbrecher, P. The QCD crossover at zero and non-zero baryon densities from lattice QCD. Preprint at <https://arxiv.org/abs/1807.05607> (2018).

Acknowledgements K.R. acknowledges support by the Polish National Science Centre under Maestro grant DEC-2013/10/A/ST2/00106. This work is part of and supported by the DFG Collaborative Research Centre ‘SFB1225/ISOQUANT’.

Author contributions All authors contributed equally to the physics analysis and to writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to P.B.-M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A homing system targets therapeutic T cells to brain cancer

Heba Samaha^{1,2,3,4}, Antonella Pignata^{2,3,4}, Kristen Fousek^{2,3,4,5}, Jun Ren⁶, Fong W. Lam^{4,7,8}, Fabio Stossi^{4,9}, Julien Dubrulle^{4,9}, Vita S. Salsman^{2,3,4}, Shanmugarajan Krishnan⁶, Sung-Ha Hong¹⁰, Matthew L. Baker^{4,11}, Ankita Shree^{2,3,4}, Ahmed Z. Gad^{1,2,3,4,5}, Thomas Shum^{2,4,5}, Dai Fukumura⁶, Tiara T. Byrd^{2,3,4,5}, Malini Mukherjee^{3,4,12}, Sean P. Marrelli¹⁰, Jordan S. Orange^{4,7,12}, Sujith K. Joseph^{2,3,4}, Poul H. Sorensen¹³, Michael D. Taylor¹⁴, Meenakshi Hegde^{2,3,4,15,16}, Maksim Mamontkin^{2,3,4,5,17}, Rakesh K. Jain⁶, Shahenda El-Naggar¹ & Nabil Ahmed^{2,3,4,5,7,15,16,17*}

Successful T cell immunotherapy for brain cancer requires that the T cells can access tumour tissues, but this has been difficult to achieve. Here we show that, in contrast to inflammatory brain diseases such as multiple sclerosis, where endothelial cells upregulate ICAM1 and VCAM1 to guide the extravasation of pro-inflammatory cells, cancer endothelium downregulates these molecules to evade immune recognition. By contrast, we found that cancer endothelium upregulates activated leukocyte cell adhesion molecule (ALCAM), which allowed us to overcome this immune-evasion mechanism by creating an ALCAM-restricted homing system (HS). We re-engineered the natural ligand of ALCAM, CD6, in a manner that triggers initial anchorage of T cells to ALCAM and conditionally mediates a secondary wave of adhesion by sensitizing T cells to low-level ICAM1 on the cancer endothelium, thereby creating the adhesion forces necessary to capture T cells from the bloodstream. Cytotoxic HS T cells robustly infiltrated brain cancers after intravenous injection and exhibited potent antitumour activity. We have therefore developed a molecule that targets the delivery of T cells to brain cancer.

The trafficking of leukocytes from the bloodstream to the brain relies on coordinated, complementary waves of expression of cell adhesion molecules (CAMs) on endothelial cells, the initial access point through the blood–brain barrier (BBB)^{1,2}. This dynamic state becomes heightened in brain infiltrative conditions, such as multiple sclerosis in which preferential access is granted to disease-mediating immune cells^{3,4}. Conversely, under the influence of cancer, homing of cytotoxic T cells is often reduced^{5,6}.

Activated leukocyte cell adhesion molecule (ALCAM; also known as CD166), a tissue-restricted CAM, has a major role in triggering T cell infiltration in inflammatory brain diseases^{7,8}. Indeed, antibodies that block ALCAM or its T cell cognate ligand, CD6, decrease leukocyte access to the brain and are in clinical trials as treatments for multiple sclerosis, HIV encephalitis and graft-versus-host disease^{9–11}. After ALCAM is engaged by T cells, successful transendothelial migration (TEM) requires that T cells sense a secondary wave of more ubiquitous CAMs on endothelial cells, predominantly mediated by ICAM1 and VCAM1, to reach the adhesion threshold needed for capture of T cells from the bloodstream¹².

We found that, similar to multiple sclerosis, brain cancer endothelial cells overexpress ALCAM; however, they downregulate ICAM1 and eliminate VCAM1, and this is likely to abrogate the homing of anti-tumour T cells. Although ALCAM is widely expressed on cancer cells and has been established as a mediator of tumour invasion and metastasis, its role in tumour endothelial cells has yet to be defined¹³. We reasoned that lessons learnt from multiple sclerosis could perhaps give us

insight into how to overcome this cancer immune-evasion mechanism; specifically, how to enable therapeutic T cells to infiltrate brain cancers.

T cell immunotherapy is an emerging field that has shown promise in clinical trials for cancer, infection, and autoimmune disease^{14,15}. Cell engineering has extended interest in this therapeutic modality; however, effective homing of therapeutic T cells to the target site remains a major limiting factor, especially for brain tumours. As cancer endothelial cells express high levels of ALCAM, but its cognate ligand CD6, which is naturally expressed on T cells, fails to mediate adequate TEM, we hypothesized that optimizing ALCAM binding by rationally re-engineering CD6 could provide an entry point for T cells through the otherwise restrictive tumour endothelium.

Cancer endothelium diverts T cells from brain tumours

We studied ALCAM expression in glioblastoma (GBM) and medulloblastoma, the most common brain cancers in adults and children, respectively, and detected intense ALCAM immunoreactivity that co-localized with CD31, denoting its vascular expression (Fig. 1a–c, Extended Data Fig. 1a). ALCAM was overexpressed on the surface of primary tumour endothelial cells (pTECs; Extended Data Fig. 1b) isolated from surgically resected GBM, in contrast to a panel of non-tumour endothelial cells, in which ALCAM was detected only intracellularly (Extended Data Fig. 2a). Notably, GBM supernatant or TGFβ¹⁶, which is highly abundant in brain cancer¹⁷, promoted ALCAM expression on endothelial cells, indicating that ALCAM is readily inducible by tumour-derived factors (Fig. 1d, Extended Data Fig. 2b).

¹Children's Cancer Hospital Egypt-57357, Cairo, Egypt. ²Center for Cell and Gene Therapy, Texas Children's Hospital, Houston Methodist Hospital and Baylor College of Medicine, Houston, TX, USA. ³Texas Children's Hospital, Houston, TX, USA. ⁴Baylor College of Medicine, Houston, TX, USA. ⁵Interdepartmental Program in Translational Biology and Molecular Medicine, Baylor College of Medicine, Houston, TX, USA. ⁶Edwin L. Steele Laboratories for Tumor Biology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. ⁷Department of Pediatrics, Baylor College of Medicine, Houston, TX, USA. ⁸Center for Translational Research on Inflammatory Diseases at the Michael E. DeBakey Veterans Affairs Medical Center, Houston, Texas, USA. ⁹Integrated Microscopy Core, Advanced Technology Cores, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA. ¹⁰Department of Neurology, McGovern Medical School at UT Health, Houston, TX, USA. ¹¹National Center for Macromolecular Imaging, Baylor College of Medicine, Houston, TX, USA. ¹²Center for Human Immunobiology, Texas Children's Hospital, Baylor College of Medicine, Houston, TX, USA. ¹³Department of Pathology & Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. ¹⁴Developmental and Stem Cell Biology Program, The Arthur and Sonia Labatt Brain Tumour Research Centre, Division of Neurosurgery, Departments of Surgery, Laboratory Medicine and Pathobiology, and of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. ¹⁵Houston Methodist Hospital, Houston, TX, USA. ¹⁶Texas Children's Cancer and Hematology Centers, Texas Children's Hospital, Baylor College of Medicine, Houston, TX, USA. ¹⁷Department of Pathology and Immunology, Baylor College of Medicine, Houston, TX, USA. *e-mail: nahmed@bcm.edu

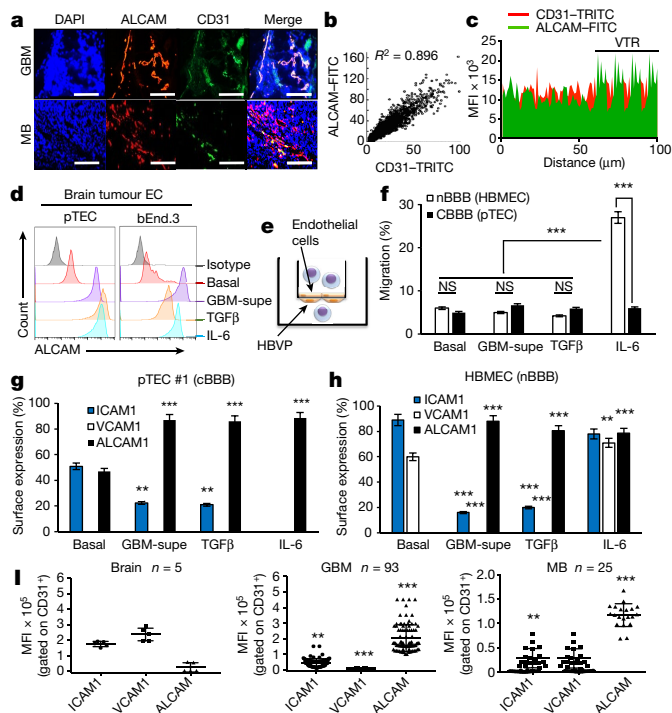


Fig. 1 | CAM expression and permeability of cancerous endothelium. **a**, Representative confocal co-immunofluorescence (IFC) of ALCAM and CD31 in 93 GBM and 25 medulloblastoma (MB), performed twice with similar results. Nuclei counterstained with DAPI. Scale bars, 100 μ m. **b**, Pearson correlation of CD31-TRITC and ALCAM-FITC pixel-mean fluorescence intensity (MFI). **c**, Topographic co-localization of CD31 and ALCAM over vascular segments (15 high-power fields per tumour averaged; representative from $n = 3$ with similar results). VTR, validation tandem-repeat. **d**, ALCAM expression in human GBM pTECs (representative of $n = 5$) and mouse brain tumour endothelium (bEnd.3) at baseline and after conditioning. Supe, supernatant. **e**, BBB model. HBVP, human brain vascular pericytes. **f**, Transmigration of T cells through BBB model. Data shown as mean \pm s.d.; Student's t -test and one-way ANOVA with Tukey's correction. *** $P < 0.001$; NS, not significant. nBBB, normal BBB. cBBB, cancer BBB. All experiments done using human T cells; validated for three donors in three or more independent experiments. **g, h**, CAM expression in pTEC#1 (**g**; $n = 5$ pTECs) and HBMEC (**h**) at baseline and after conditioning. **i**, High-throughput CAM quantification in five normal brains, 93 GBM and 25 medulloblastoma, each examined twice. Each point is an average of MFI acquired from 15 confocal CD31⁺-gated vascular-patterned high-power fields and segmented by channel-specific intensity thresholding per tumour. **g–i**, Data shown as mean \pm s.d.; ANOVA with Tukey's correction; ** $P < 0.01$, *** $P < 0.001$.

During T cell migration across the BBB in multiple sclerosis, endothelial ALCAM co-localizes in lipid rafts with T cell CD6^{18,19}. To investigate whether the observed ALCAM overexpression could enable T cell transmigration through a cancer BBB, we created an in vitro BBB model by sandwiching a polycarbonate membrane between pTECs and pericytes (Fig. 1e). Despite ALCAM overexpression (Extended Data Fig. 2c), the cancer BBB remained impermeable to T cells both at baseline and after conditioning with GBM supernatant or TGF β (Fig. 1f). By contrast, conditioning of a normal BBB, in which pTECs were replaced with normal brain endothelial cells, with IL-6 rendered it highly permissive. The fact that cancer BBB was resistant to the effects of pro-inflammatory and tumour stroma-secreted IL-6 suggested that ALCAM overexpression alone is insufficient to enable transmigration of T cells, indicating that perhaps the secondary wave of adhesion, well described in multiple sclerosis, is lacking¹².

We therefore studied the dynamic expression of the principal mediators of the secondary wave on cancer endothelial cells. In contrast to normal brain endothelial cells, we found that pTECs express lower levels of ICAM1 and no VCAM1 at baseline (Fig. 1g, h, Extended

Data Fig. 2d–h). Culturing pTECs in the presence of GBM supernatant, TGF β or IL-6 further decreased ICAM1 and markedly upregulated ALCAM. As expected, IL-6 increased the expression of ICAM1, VCAM1 and ALCAM in normal brain endothelial cells. Notably, we observed this distinct pattern of adhesion molecule expression in the microvasculature of surgically excised GBM ($n = 93$) and medulloblastoma ($n = 25$) but not in normal brains ($n = 5$; Fig. 1i, Extended Data Fig. 1a).

The reduction of ICAM1 and elimination of VCAM1 under the influence of tumour-derived factors suggested that pTECs have inherent resistance to interactions with T cells, which explained the impermeability of the cancer BBB seen in Fig. 1f.

Engineering T cells to traverse cancer endothelium

ALCAM expression was intensified in cancer endothelial cells; we therefore reasoned that enhanced T cell anchorage to ALCAM could compensate for the reduced expression of ICAM1. As constitutively expressed CD6 on T cells was insufficient to promote effective TEM, we rationally re-engineered CD6, to create a system for guided homing of T cells to brain cancers.

To extract the homing function of CD6, we computationally mapped the ALCAM binding region to extracellular domain 3 (D3) of CD6, in agreement with previous reports^{9,10} (Fig. 2a, Extended Data Fig. 3a–c). The prototype HS molecule included a D3 exodomain, an IgG1 hinge and a CD6 transmembrane and signalling domains (Fig. 2b, Extended Data Fig. 3d). We multimerized the exodomain (creating a trimer (3HS) and a pentamer (5HS)) to study the effect of enhancing the molecule's crosslinking avidity to ALCAM on T cell behaviour (Fig. 2c, Extended Data Fig. 3e) and created tail-less HS Δ molecules to study the role of HS signalling (Fig. 2d, Extended Data Fig. 3f). We detected HS molecules on the surface of human T cells by probing D3 (Extended Data Fig. 3g, h) and confirmed that they bound specifically to ALCAM (Fig. 2e, Extended Data Fig. 3g). The molecular interactions of T cells with the endothelium at the trans migratory cup are mediated by podosynaptic CAMs engaging their cognate ligands to initiate TEM²⁰. We detected D3–ALCAM heterodimers in HS T cell trans migratory cups; these intensified with D3 multimerization²⁰ (Fig. 2f, Extended Data Fig. 3i, j).

Capture of T cells from the bloodstream, rolling along the vessel wall and firm adhesion are key steps before TEM²¹. To study the effect of HS on TEM kinetics, we used microfluidics flow chambers lined by TGF β -conditioned ALCAM⁺ endothelium and applied 2 dyne cm⁻² shear force, akin to that found in tumour capillary vessels (Extended Data Fig. 3k, l, Supplementary Video 1). HS T cells were captured more frequently, rolled more slowly, and stopped more quickly than normal T cells (Fig. 2g–i, Extended Data Fig. 3m, n). Multimerization of D3 increased the capture and arrest of T cells, while inclusion of a signalling domain maximized this effect independent of the exodomain. Rolling velocities were similar between T cells expressing different versions of HS, probably because rolling is predominantly influenced by selectins, which are expressed similarly on all T cells²². Before extravasation, HS T cells were more resistant to mechanical detachment (Fig. 2j) than normal T cells and could therefore resist fluctuations in blood flow typical of tumour neovasculature. HS T cells did not engage normal (ALCAM-negative) endothelial cells.

Next, we evaluated the ability of HS to enable T cell transmigration through our BBB model (Fig. 2k). Resting BBB allowed negligible T cell migration, but induction of ALCAM led to a substantial increase in migration of HS T cells. Multimerization of the exodomain further increased this migration (5HS $\Delta > 3HS\Delta > HS\Delta$) and a signalling domain maximized migration across constructs. Blocking the D3–ALCAM interaction using soluble ALCAM abrogated migration with all HS T cells, and washing it largely restored the pre-blocking pattern. Molecular interruption of ALCAM expression on pTECs using small interfering RNA (siRNA) and on human umbilical vein endothelial cells (HUVECs) using CRISPR–Cas9 muted their responsiveness to TGF β (Extended Data Fig. 4a–f) and reduced the transmigration of HS T cells

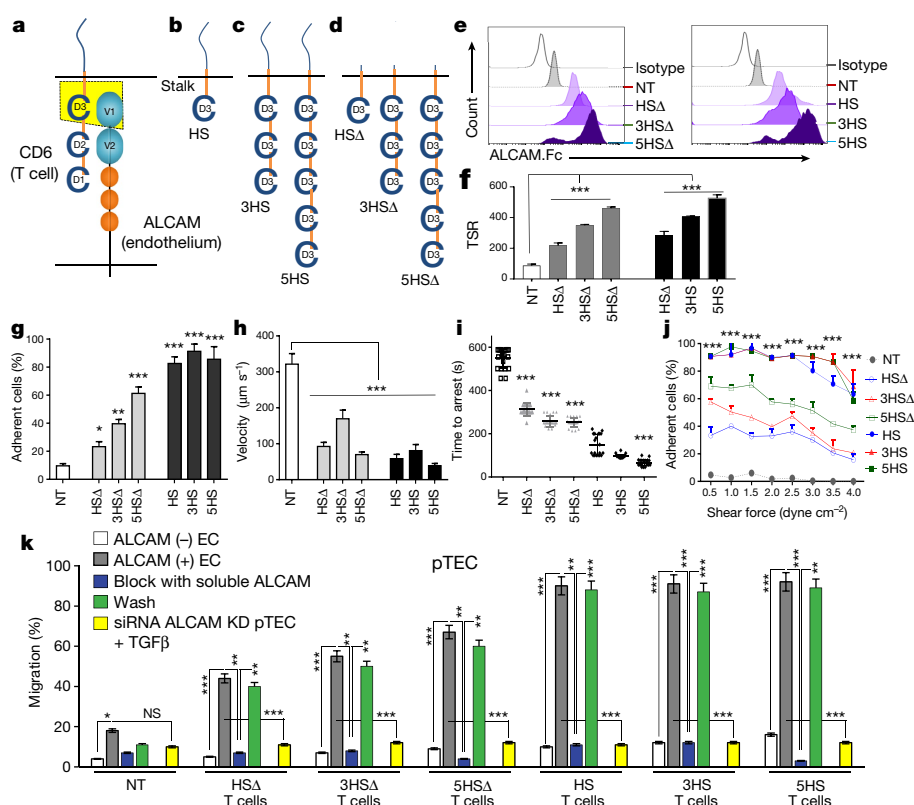


Fig. 2 | Rational engineering of the homing system. **a**, Schematic of ALCAM binding region on CD6. **b**, Prototype HS molecule. **c**, Multimerized exodomains. **d**, Tail-less HSA Δ molecules lacking signalling domains. **e**, Specific binding to soluble ALCAM; assessed independently ten times with similar results. NT, normal T cells. **f**, Proximity ligation assay (PLA) identifying D3-ALCAM heterodimers in conjugates of T cells and endothelium. 83–103 cell conjugates analysed per condition; repeated three times independently with similar results. TSR, total signals per region. **g–j**, TEM kinetics of 1×10^6 cells per

condition in microfluidic channels under $1\text{--}3\text{ dyne cm}^{-2}$ shear over ALCAM $^+$ pTECs. **k**, Transmigration of 2×10^5 T cells per well through pTEC cBBB and the effect of soluble ALCAM blockade and washing. Yellow bars show the effect of ALCAM siRNA knockdown on the permissivity of endothelial cells (EC). Three experiments independently performed in triplicate with similar results. All data shown as mean \pm s.d. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. One-way ANOVA then Dunnett's test for multiple comparisons (compared to normal T cells).

(yellow bars in Fig. 2k). Homotypic ALCAM-ALCAM binding occurs but is decimated by the 50-fold stronger heterotypic ALCAM-CD6 interaction¹. We found that elimination of T cell ALCAM improved the transmigration of HSA Δ (but not of HS) T cells, indicating that the transmigration of HS T cells is largely mediated through heterotypic interactions of HS with endothelial ALCAM¹ (Extended Data Fig. 4g–j).

We concluded that enhanced ALCAM anchorage can trigger BBB transmigration of HS T cells. Multimerization of the HS exodomain produced an incremental increase in TEM; however, signalling through the HS endodomain maximized this effect, indicating that other mechanistic downstream events might be involved in enabling transmigration of HS T cells.

HS harnesses the adhesive power of ICAM1-LFA-1 axis

Although the definite functions of CD6 remain unclear, its signalling is thought to be critical for T cell motility and cell-cell contact¹¹. CD6 activates the downstream adaptor SH2-domain containing leukocyte protein of 76 kD (SLP-76) in a shared pathway that results in activation of lymphocyte function-associated antigen-1 (LFA-1). Upon activation, LFA-1 undergoes a conformational change that exposes a ligand binding site for ICAM1^{23–25}. We reasoned that if HS signalling converges on SLP-76, ICAM1 could be brought into play, explaining the functional superiority of HS T cells over HSA Δ T cells.

We sequentially interrogated individual mediators downstream of the HS endodomain that could link to integrin modulation. We found substantial co-clustering of SLP-76 with eGFP-tagged HS molecules (HS-eGFP) but not with eGFP-tagged HSA Δ -molecules (HSA Δ -eGFP) upon crosslinking with ALCAM immobilized on a glass surface

(Fig. 3a). SLP-76 binds to CD6 in a zeta-chain-associated protein kinase-70 (Zap-70)-dependent manner, and, in contrast to HSA Δ and normal T cells, all HS T cells showed higher ZAP-70 phosphorylation after transmigration through the cancer BBB model²⁶. HS T cells also recruited cytosolic talin, a high-molecular-weight cytoskeletal protein that unfolds LFA-1 into a high-affinity conformation^{27,28}. Indeed, upon transmigration, talin co-aggregated with and unfolded LFA-1 and engaged ICAM1 at trans migratory cups formed by HS T cells moving between and through ALCAM-expressing endothelial cells (Fig. 3b, c, Extended Data Fig. 5a–c). HSA Δ T cells did not show these effects.

We functionally confirmed this finding: BIRT 377, an allosteric inhibitor of unfolded LFA-1²⁹, substantially reduced the migratory capacity of HS T cells to levels comparable to those of HSA Δ T cells, highlighting the contribution of the CD6 signalling endodomain (Fig. 3d). The effect of BIRT 377 on the migratory capacity of HSA Δ T cells and normal T cells was negligible.

Thus, our data confirmed that CD6 signalling culminates in the unfolding of LFA-1 on T cells, enabling HS T cells to engage low-level ICAM1 and providing the deficient secondary wave of adhesion to cancer endothelial cells (Fig. 3e).

Cytoskeletal changes mediate adept transmigration

Talin is concentrated at regions of cell-cell contact, and drives the extravasation of T cells through cortical actin polymerization and maturation of focal adhesion complexes mediated by focal adhesion kinase (FAK) and vinculin^{30,31}. We used total internal reflection fluorescence (TIRF) and found that upon landing of T cells on ALCAM-coated glass surfaces, actin was induced and co-localized with HS-eGFP but

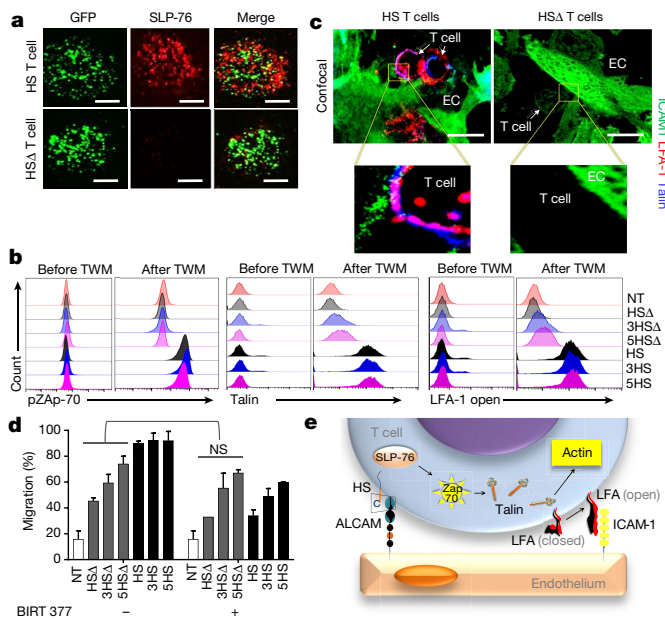


Fig. 3 | Signalling events downstream of HS molecules. **a**, Confocal IFC images after 5×10^4 T cells landed on an ALCAM-coated glass surface, showing micro-clusters of SLP-76 (red) and eGFP-tagged (green) HS and HSΔ T cells. Scale bars, 50 μm. **b**, Intracellular flow cytometry for pZap70, talin and surface staining for unfolded LFA-1 using KIM127, monoclonal antibodies that bind exclusively to the extended $\beta 2$ -chain (CD18), before and after transmigration of 2×10^5 T cells through an ALCAM⁺ cBBB model. **c**, Confocal images of the transmigration cup at the HS T cell-endothelial cell interface co-stained for talin (blue), ICAM1 (green) and unfolded LFA-1 (red). Scale bars, 10 μm. **d**, T cell transmigration through cBBB model before and after incubation with the LFA-1 allosteric inhibitor, BIRT 377. Data shown as mean \pm s.d., four independent experiments with similar results. $P = 0.774$. Mann-Whitney U -test; NS, not significant. **e**, Schematic of HS signalling events culminating in unfolding of LFA-1.

not with HSΔ-eGFP on T cells or with normal T cells (Fig. 4a, b). Signalling from FAK promotes adhesion maturation of the migrating T cells and mediates the rear retraction of T cells crawling on endothelial cells and their ultimate protrusion and extravasation^{32,33}. We found denser actin and FAK after ALCAM interaction in HS T cells than in HSΔ T cells or normal T cells (Fig. 4c, d). We also detected membrane ruffling, the formation of a motile cell surface containing a meshwork of newly polymerized actin, and an enrichment of actin-FAK in lamellipodia and invadopodia using structure illuminating microscopy (SIM), which offers higher lateral resolution (Fig. 4e).

To quantify these findings in a technically unbiased manner, we performed high-throughput deconvolution microscopy on T cells from three donors at different levels of ALCAM (Fig. 4f, g). HS T cells expressed more actin and FAK than normal T cells. In addition, actin and FAK co-localized at the surface, enabling protrusion of the podosynaptic structure of the T cells, which is needed for subsequent endothelial invasion. All HS T cells had more podosynaptic lamellipodia and focal adhesions per cell, and a significantly larger area of spread, than normal T cells, meaning that the cytoskeletal rearrangement is well-tensioned to enable T cell migration.

Collectively, subcellular microscopy demonstrated that HS molecules anchor to the actin cytoskeleton and induce maturation of FAK to enable T cell transmigration.

Targeted homing of HS T cells to brain cancer

ALCAM and its binding region on CD6 are highly conserved, with 93–96% homology between human and mouse, enabling us to assess the ability of intravenous HS T cells to overcome the endothelial blockade and home to orthotopic U87-GBM in severe combined immunodeficiency (SCID) mice (Fig. 5a). Flow cytometry

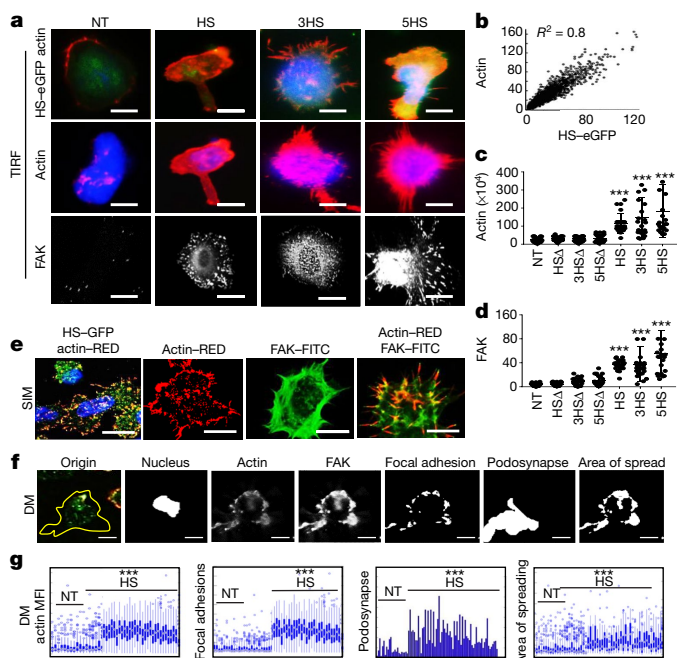


Fig. 4 | Cytoskeletal changes mediated by HS signalling.

a, Representative TIRF micrographs of HS-eGFP T cells upon landing on an ALCAM-coated glass surface. Scale bars, 10 μm. **b**, HS-eGFP as in **a** correlated with actin. **c**, **d**, Comparison of actin (**c**) and FAK MFI (**d**) among HS and HSΔ T cells. **e**, Representative SIM images depicting HS T cell membrane ruffles. Scale bars, 10 μm. **f**, Use of MATLAB script to analyse deconvolution microscopy (DM) data of podosynaptic protrusions and their spread in 2×10^6 HS T cells. Scale bars, 10 μm. **g**, Characterization of migrating T cells through collective quantification of actin MFI, focal adhesions, area of spreading, and podosynapse formation by high-throughput deconvolution microscopy at HS-ALCAM interface in a representative donor ($n = 200$ –800 cells per condition). Each column represents cells in one well. All assessments independently repeated three times with similar results. Data shown as mean \pm s.d., *** $P < 0.001$. Tukey's test used in **c**, **d**; Student's t -test used in **g**.

of tumour-infiltrating lymphocytes (TILs) demonstrated that all HS T cells had superior specific homing capacity compared to normal T cells, and that 5HS T cells had the densest TIL infiltrate (Fig. 5b, Extended Data Fig. 6a).

Next, we injected eGFP–firefly-luciferase (eGFP–FFLuc)-labelled T cells intravenously into mice harbouring U87-GBM (adult GBM) and Daoy-MB (paediatric medulloblastoma) orthotopic grafts, and quantified T cell homing using bioluminescence imaging (BLI). HS T cells had a 1–2 log brighter signal than normal T cells in U87-GBM (Fig. 5c, d) and Daoy-MB (Fig. 5e, f). Analysis of the spatial orientation of T cells to the 3D reconstituted tumour vasculature in tumour explants showed substantially higher HS T cell signals in the intra- and perivascular areas compared to normal T cells (Fig. 5g, h). Finally, we used a cranial window and video-rate multiphoton microscopy to examine the in vivo dynamics of HS T cell homing to U87-GBM with single-cell resolution (Fig. 5i, Extended Data Fig. 6b). Continuous videography of size-matched vasculature showed more rolling 5HS T cells than normal T cells along tumour vessels and achieving firm arrest (Fig. 5j, k, Supplementary Videos 2–5). Reconstitution of time-lapse images in 3D showed more extravasating HS T cells than normal T cells at the tumour vascular bed (Fig. 5l).

We investigated whether intravenous 5HS T cells invaded normal tissues, but found negligible TILs in the spleen, lungs and kidneys of U87-GBM-bearing mice, at levels no greater than for normal T cells (Extended Data Fig. 7a). We also observed no T cells or alterations in the histomorphology of normal brain tissue, despite the presence of a heavy HS T cell infiltrate in the tumour (Extended Data Fig. 7b).

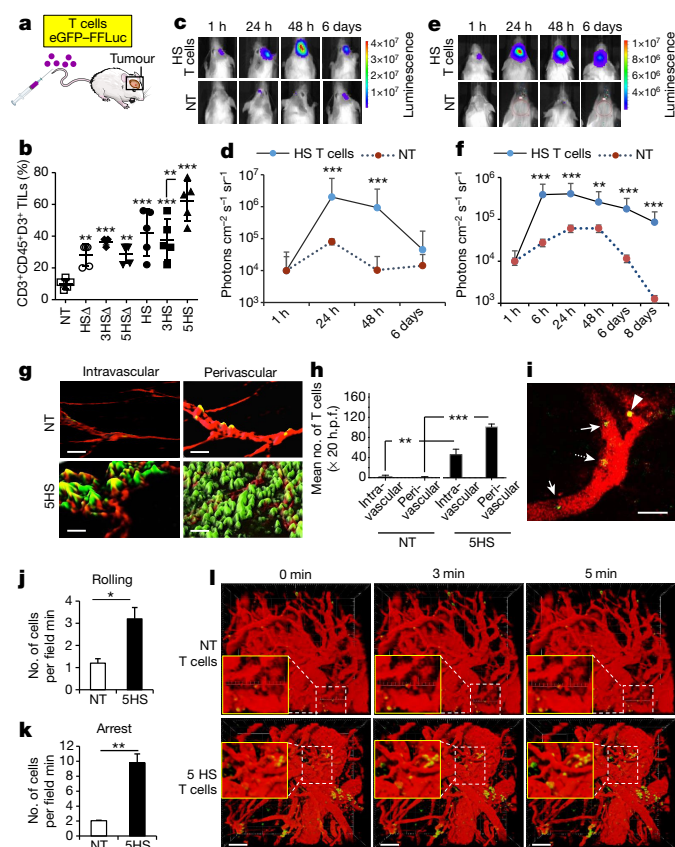


Fig. 5 | Homing of HS T cells to brain tumours. **a**, eGFP-labelled T cells were injected intravenously into mice bearing orthotopic tumours. **b**, Flow cytometry analysing TILs in GBM explants ($n = 5$ mice per group). **c–f**, BLI of T cells after intravenous injection into GBM (**c**, quantified in **d**) and medulloblastoma (**e**, quantified in **f**). Data shown as mean \pm s.d. ($n = 5$ mice per group). $^{**}P = 0.001$, $^{***}P < 0.0001$. ANOVA with Tukey's correction. **g**, Iso-surface 3D rendering of tumour explant confocal images showing eGFP-HS T cells (green) relative to ALCAM⁺ vessels (red). Cryo-sections imaged at $40\times$, $50\mu\text{m}$ z-stacks, scale bars, $50\mu\text{m}$. **h**, Quantification of GFP⁺ T cells in and around the ALCAM⁺ (red) signal indicating perivascular and intravascular locations, respectively. Data shown as mean \pm s.d. ($n = 4$ explants). $^{**}P = 0.0015$, $^{***}P < 0.0001$. Two-tailed t -test. **i–k**, Dynamics of T cell homing. **i**, Snapshot image taken at 15 s of Supplementary Video M2 showing rolling (arrow) and adherent (dashed arrow) 5HS T cells inside the blood vessels, and a 5HS T cell extravasating (arrowhead) into a U87-GBM tumour. Green, T cells; red, TAMRA-dextran (blood vessels). Scale bar, $100\mu\text{m}$. **j**, **k**, Quantification of rolling (**j**) or adherent (**k**) T cells in U87-GBM vasculature. $n = 3$ mice per group. Data shown as mean \pm s.e.m. $^{*}P < 0.05$, two-tailed t -test. $^{*}P = 0.0219$ (**j**) and $^{**}P = 0.0033$ (**k**). **l**, Time-lapse 3D reconstructed images showing extravasation of T cells. Green, T cells; red, TAMRA-dextran (blood vessels). Scale bars, $100\mu\text{m}$.

Anti-tumour activity of cytotoxic HS T cells

Next, we assessed the ability of HS T cells to deliver a therapeutic complex biological agent to brain cancers (Fig. 6a). We armed the most successful design, 5HS T cells, with chimaeric antigen receptors (CARs) specific for human epidermal growth factor receptor 2 (HER2), a glioma antigen currently targeted by CAR T cells in several clinical trials³⁴ (Extended Data Fig. 8a). Before testing in animals, we confirmed that only HER2-CAR 5HS T cells, but not 5HS T cells or normal T cells, efficiently killed U87-GBM cells in vitro (Fig. 6b). Notably, HER2-CAR 5HS T cells did not lyse normal or tumorous human or mouse endothelial cells (Fig. 6c) or any other ALCAM-expressing leukocytes (Fig. 6d). Thus, HS T cells have no cytolytic activity against ALCAM-expressing targets, and cytolysis is distinctly mediated through engagement of HER2 by the CAR molecules. Similar to

HER2-CAR T cells, all HER2-CAR HS T cells exhibited a predominantly effector-memory phenotype and comparable exhaustion and proliferation profiles following transwell migration (TWM; Extended Data Fig. 8b–d).

To test the antitumour efficacy of HER2-CAR 5HS T cells, we established eGFP-FFLuc-labelled orthotopic U87-GBM tumours. Unlabelled T cells were injected intravenously, and tumour growth was monitored. HER2-CAR 5HS T cells induced regression of established tumours in all treated animals, in contrast to HER2-CAR T cells, which transiently slowed tumour growth, and normal T cells (Fig. 6e, f). HER2-CAR 5HS T cells were more abundant than HER2-CAR T cells and normal T cells in GBM explants (Fig. 6g, Extended Data Fig. 9a) but were absent from non-tumour areas of the brain (Extended Data Fig. 9b). Kaplan–Meier survival analysis showed that mice treated with HER2-CAR 5HS T cells had a median survival exceeding 60 days, compared to 22 and 18 days for animals treated with HER2-CAR T cells and normal T cells, respectively (Fig. 6h).

Discussion

The onset of TEM is mediated by waves of engagement of endothelial cell adhesion molecules to their cognate ligands on T cells, signalling by which mediates the adhesion threshold necessary to pull T cells from the bloodstream³⁵. In multiple sclerosis, this threshold is reached by ALCAM crosslinking CD6, followed by a secondary wave of interaction with other ubiquitous CAMs, predominantly ICAM1³⁶.

We have shown that, similar to inflammatory endothelial cells, cancer endothelial cells overexpress ALCAM, but abrogate T cell homing by downregulating ICAM1. Our results indicate that the interaction between endogenous CD6 on T cells and ALCAM on cancer endothelial cells alone is incapable of mediating T cell transmigration. We thus created HS, a set of engineered ligands, to enhance the transmigration of T cells through cancer endothelial cells. We demonstrate how HS T cells can harness the power of a preexisting pathway (ICAM1–LFA-1), akin to the secondary wave seen in multiple sclerosis^{37,38}. This transforms the ineffective T cell–cancer endothelial cell interaction into a permissive inflammatory one. Subsequently, T cells are guided by a chemokine gradient to their tumour target (Fig. 6i). It is likely that this chemokine gradient could interact with and influence the process of TEM.

We created HS to be an abstract ligand based on the minimal moiety on CD6 that could heterodimerize with ALCAM. In doing so, we extracted the homing function of CD6 (mediated by D3) and avoided carrying forward its other unwanted functions, which remain mostly unclear¹¹. Indeed, when we cloned and expressed the full-length CD6 on human T cells, their transmigration was improved but they were exceedingly activated at baseline and exhibited a rather exhausted phenotype, failing to expand (Extended Data Fig. 10). This confirms recent literature describing D1 of CD6 as a mediator for T cell activation³⁹. It also underscores the critical role of rational engineering in the design of synthetic molecules to optimally mediate distinct functions, while simultaneously avoiding unwanted ones.

Poor homing of CAR T cells is a major obstacle to effective T cell therapy^{40,41}. We recently established a favourable safety profile of intravenous HER2-CAR T cells in patients with GBM but observed only disease stabilization and partial responses⁴². As an alternate strategy, we and others have reverted to intra-tumoral administration to achieve the bioavailability needed to elicit a more uniform clinical response (NCT02442297); this approach is rather invasive and of limited applicability. At this stage, we ought to seek highly specific refinements to cellular therapeutics, such as the HS platform, if more durable tumour clinical responses are to be achieved.

HS molecules enhance the ability of therapeutic T cells to exert an antitumour activity while maintaining a favourable safety profile. Adaptations to the HS platform as a modular delivery tool could be made to allow complex therapeutic agents to access diseased brain sites.

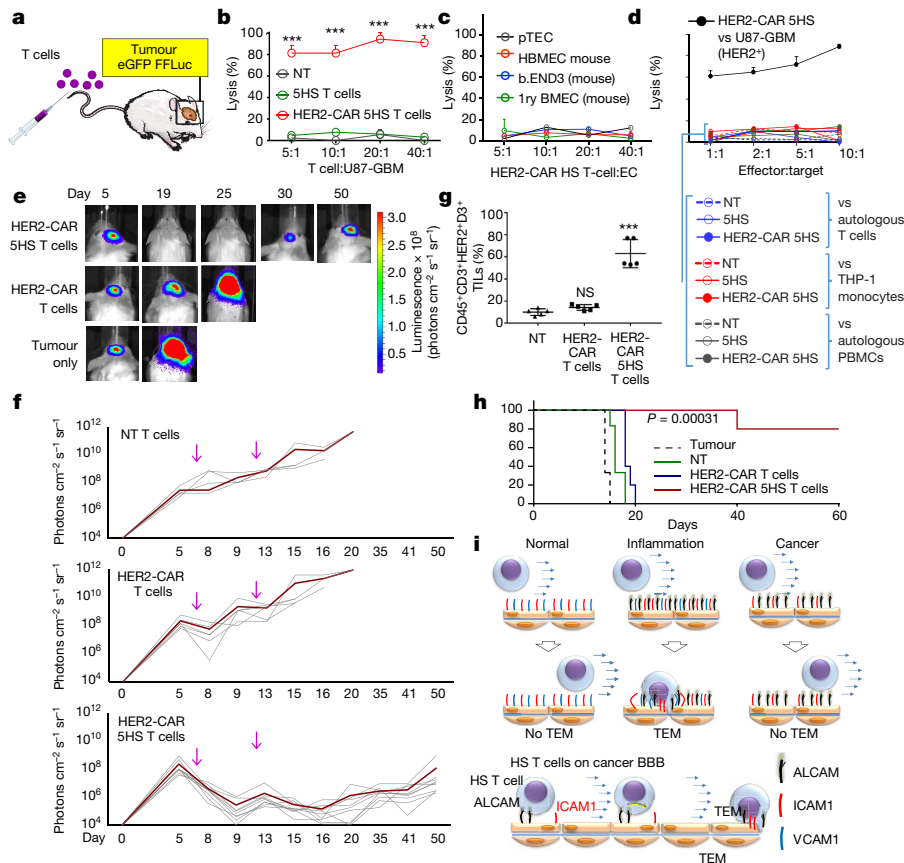


Fig. 6 | Anti-tumour activity of cytotoxic HS T cells. a, Schematic of experiment. **b–d**, ^{51}Cr -cytotoxicity assessing the cytolytic activity of HS T cells at indicated effector:target ratios against 5×10^3 targets. **b**, HER2⁺ U87-GBM; **c**, human and mouse endothelial cells; **d**, ALCAM-expressing leukocytes. THP-1, human monocytic cells. PMBC, peripheral blood mononuclear cells. Data shown as mean of triplicate \pm s.d.; *** $P < 0.001$, one-way ANOVA with post-hoc Tukey's test. Three experiments from three donors done with similar results. **e**, BLI of tumours ($n = 5$ –10 mice

per group) after intravenous injection of T cells; quantified in **f** (arrows indicate T cell injection). **g**, Flow cytometry quantifying TILs in explants. Data shown as mean \pm s.d. Four experiments done with similar results, $P < 0.001$. Tukey's test. **h**, Kaplan–Meier survival probability analysed by log-rank test, *** $P = 0.00034$. **i**, Schematic summarizing how the HS platform transforms the obstructive cancer endothelium into a selectively permissive inflammatory-like one, allowing enhanced targeted delivery of T cells.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0499-y>

Received: 11 October 2017; Accepted: 8 August 2018;

Published online: 05 September 2018

- Davenport, M. P., Grimm, M. C. & Lloyd, A. R. A homing selection hypothesis for T-cell trafficking. *Immunol. Today* **21**, 315–317 (2000).
- Krummel, M. F., Bartumeus, F. & Gérard, A. T cell migration, search strategies and mechanisms. *Nat. Rev. Immunol.* **16**, 193–201 (2016).
- Carrithers, M. D., Visintin, I., Kang, S. J. & Janeway, C. A., Jr. Differential adhesion molecule requirements for immune surveillance and inflammatory recruitment. *Brain* **123**, 1092–1101 (2000).
- Arima, Y. *et al.* Regulation of immune cell infiltration into the CNS by regional neural inputs explained by the gate theory. *Mediators Inflamm.* **2013**, 898165 (2013).
- Sackstein, R., Schattton, T. & Barthel, S. R. T-lymphocyte homing: an underappreciated yet critical hurdle for successful cancer immunotherapy. *Lab. Invest.* **97**, 669–697 (2017).
- Fukumura, D., Kloepper, J., Amoozgar, Z., Duda, D. G. & Jain, R. K. Enhancing cancer immunotherapy using antiangiogenics: opportunities and challenges. *Nat. Rev. Clin. Oncol.* **15**, 325–340 (2018).
- Cayrol, R. *et al.* Activated leukocyte cell adhesion molecule promotes leukocyte trafficking into the central nervous system. *Nat. Immunol.* **9**, 137–145 (2008).
- Nelissen, J. M. D. T., Peters, I. M., de Grooth, B. G., van Kooyk, Y. & Figdor, C. G. Dynamic regulation of activated leukocyte cell adhesion molecule-mediated homotypic cell adhesion through the actin cytoskeleton. *Mol. Biol. Cell* **11**, 2057–2068 (2000).
- Brown, M. H. CD6 as a cell surface receptor and as a target for regulating immune responses. *Curr. Drug Targets* **17**, 619–629 (2016).
- Chappell, P. E. *et al.* Structures of CD6 and its ligand CD166 give insight into their interaction. *Structure* **23**, 1426–1436 (2015).

- Li, Y. *et al.* CD6 as a potential target for treating multiple sclerosis. *Proc. Natl Acad. Sci. USA* **114**, 2687–2692 (2017).
- Bullard, D. C. *et al.* Intercellular adhesion molecule-1 expression is required on multiple cell types for the development of experimental autoimmune encephalomyelitis. *J. Immunol.* **178**, 851–857 (2007).
- Kijima, N. *et al.* CD166/activated leukocyte cell adhesion molecule is expressed on glioblastoma progenitor cells and involved in the regulation of tumor cell invasion. *Neuro-oncol.* **14**, 1254–1264 (2012).
- Rosenberg, S. A. & Restifo, N. P. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* **348**, 62–68 (2015).
- Bonini, C. & Mondino, A. Adoptive T-cell therapy for cancer: The era of engineered T cells. *Eur. J. Immunol.* **45**, 2457–2469 (2015).
- Hansen, A. G. *et al.* ALCAM/CD166 is a TGF- β -responsive marker and functional regulator of prostate cancer metastasis to bone. *Cancer Res.* **74**, 1404–1415 (2014).
- Peñuelas, S. *et al.* TGF- β increases glioma-initiating cell self-renewal through the induction of LIF in human glioblastoma. *Cancer Cell* **15**, 315–327 (2009).
- Gu, M.-X. *et al.* Proteomic analysis of endothelial lipid rafts reveals a novel role of statins in antioxidant. *J. Proteome Res.* **11**, 2365–2373 (2012).
- Dorovini-Zis, K. *The Blood–Brain Barrier in Health and Disease, Volume One: Morphology, Biology and Immune Function* (CRC Press, London, 2015).
- Carman, C. V. & Springer, T. A. A transmembrane cup in leukocyte diapedesis both through individual vascular endothelial cells and between them. *J. Cell Biol.* **167**, 377–388 (2004).
- Muller, W. A. Mechanisms of leukocyte transendothelial migration. *Annu. Rev. Pathol.* **6**, 323–344 (2011).
- McEver, R. P. & Zhu, C. Rolling cell adhesion. *Annu. Rev. Cell Dev. Biol.* **26**, 363–396 (2010).
- Engelhardt, B. Molecular mechanisms involved in T cell migration across the blood-brain barrier. *J. Neural Transm. (Vienna)* **113**, 477–485 (2006).
- Laschinger, M., Vajkoczy, P. & Engelhardt, B. Encephalitogenic T cells use LFA-1 for transendothelial migration but not during capture and initial adhesion strengthening in healthy spinal cord microvessels in vivo. *Eur. J. Immunol.* **32**, 3598–3606 (2002).

25. Green, C. E. *et al.* Dynamic shifts in LFA-1 affinity regulate neutrophil rolling, arrest, and transmigration on inflamed endothelium. *Blood* **107**, 2101–2111 (2006).
 26. Orta-Mascaró, M. *et al.* CD6 modulates thymocyte selection and peripheral T cell homeostasis. *J. Exp. Med.* **213**, 1387–1397 (2016).
 27. Calderwood, D. A. & Ginsberg, M. H. Talin forges the links between integrins and actin. *Nat. Cell Biol.* **5**, 694–697 (2003).
 28. Lawson, C. *et al.* FAK promotes recruitment of talin to nascent adhesions to control cell motility. *J. Cell Biol.* **196**, 223–232 (2012).
 29. Poria, R. B. *et al.* Characterization of a radiolabeled small molecule targeting leukocyte function-associated antigen-1 expression in lymphoma and leukemia. *Cancer Biother. Radiopharm.* **21**, 418–426 (2006).
 30. Baumann, K. Cell adhesion: FAK or talin: who goes first? *Nat. Rev. Mol. Cell Biol.* **13**, 138–139 (2012).
 31. Critchley, D. R. Cytoskeletal proteins talin and vinculin in integrin-mediated adhesion. *Biochem. Soc. Trans.* **32**, 831–836 (2004).
 32. Mitra, S. K., Hanson, D. A. & Schlaepfer, D. D. Focal adhesion kinase: in command and control of cell motility. *Nat. Rev. Mol. Cell Biol.* **6**, 56–68 (2005).
 33. Cavalcanti-Adam, E. A. *et al.* Cell spreading and focal adhesion dynamics are regulated by spacing of integrin ligands. *Biophys. J.* **92**, 2964–2974 (2007).
 34. Hegde, M. *et al.* Tandem CAR T cells targeting HER2 and IL13R α 2 mitigate tumor antigen escape. *J. Clin. Invest.* **126**, 3036–3052 (2016).
 35. Liu, Y. *et al.* Regulation of leukocyte transmigration: cell surface interactions and signaling events. *J. Immunol.* **172**, 7–13 (2004).
 36. Auerbach, S. D., Yang, L. & Luscinskas, F. W. in *Adhesion Molecules: Function and Inhibition* 99–116 (Springer, Basel, 2007).
 37. Steiner, O. *et al.* Differential roles for endothelial ICAM-1, ICAM-2, and VCAM-1 in shear-resistant T cell arrest, polarization, and directed crawling on blood-brain barrier endothelium. *J. Immunol.* **185**, 4846–4855 (2010).
 38. Lee, B. P. L. & Imhof, B. A. Lymphocyte transmigration in the brain: a new way of thinking. *Nat. Immunol.* **9**, 117–118 (2008).
 39. Bughani, U. *et al.* T cell activation and differentiation is modulated by a CD6 domain 1 antibody Itolizumab. *PLoS One* **12**, e0180088 (2017).
 40. Ager, A., Watson, H. A., Wehenkel, S. C. & Mohammed, R. N. Homing to solid cancers: a vascular checkpoint in adoptive cell therapy using CAR T-cells. *Biochem. Soc. Trans.* **44**, 377–385 (2016).
 41. D'Aloia, M. M., Zizzari, I. G., Sacchetti, B., Pierelli, L. & Alimandi, M. CAR-T cells: the long and winding road to solid tumors. *Cell Death Dis.* **9**, 282 (2018).
 42. Ahmed, N. *et al.* HER2-specific chimeric antigen receptor-modified virus-specific T cells for progressive glioblastoma: a phase 1 dose-escalation trial. *JAMA Oncol.* **3**, 1094–1101 (2017).
- Acknowledgements** We thank M. K. Brenner and C. Gillespie for scientific advice and linguistic editing, respectively, and S. Roberge and M. Duquette for technical assistance. The D3 antibody was a gift from M. Brown. This work was funded by an SU2C-St. Baldrick's Pediatric Dream Team Translational Research Grant (SU2C-AACR-DT1113; NA/PS/MDT). SU2C is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. Also funded by Alex's Lemonade Stand, NIH-T32HL092332 (K.F./T.B.; H. Heslop), R01AI067946 (J.S.O.), T32GM08812 (K.F./T.B.; M. Estes), DK56338, CA125123, and 1S10OD020151-01, CPRIT (RP150578), the Dan L. Duncan CCC and P01-CA080124 (R.K.J./D.F.); R35-CA197743, K08-GM123261 (F.L.) and P50-CA165962 (R.K.J.). This content does not necessarily represent the official views of the funding agencies, the Department of Veterans Affairs or the U.S. Government.
- Reviewer information** Nature thanks M. Platten and the other anonymous reviewer(s) for their contribution to the peer review of this work.
- Author contributions** N.A. conceived the main study idea, and with H.S. conceived and implemented the study details. N.A., K.F. and A.P. designed HS molecules. H.S., M.D.T., S.P.M., P.S., S.E.-N., M.H., F.S., J.D. and N.A. performed the CAM studies. M.L.B. performed in silico modelling. F.L. and H.S. designed and implemented microfluidic experiments. M.Ma., T.B., S.K.J. and A.Z.G. performed molecular testing. H.S., F.S., J.D., M.Mu. and J.S.O. designed and performed the subcellular imaging experiments. J.R., H.S., V.S.S., A.S., T.S., S.P.M., S.-H.H., D.F., S.K., R.K.J. and N.A. implemented the animal microscopy and experiments. All authors gave final approval.
- Competing interests** The authors declare no competing interests.
- Additional information**
- Extended data** is available for this paper at <https://doi.org/10.1038/s41586-018-0499-y>.
- Supplementary information** is available for this paper at <https://doi.org/10.1038/s41586-018-0499-y>.
- Reprints and permissions information** is available at <http://www.nature.com/reprints>.
- Correspondence and requests for materials** should be addressed to N.A.
- Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Antibodies, recombinant proteins and chemicals. *Antibodies.* Anti-FAK, anti-ALCAM-PE, anti-ICAM1-PE, anti-VCAM1-PE, anti-VE cadherin, anti-von Willebrand (vWF) factor and anti-CHS1 were purchased from Abcam (Cambridge, MA, USA). OX124, mouse anti-human monoclonal antibody against HS, was a kind gift from M. Brown, Oxford, UK. Biotin-labelled anti-CD18 KIM127 was purchased from Exploratory Research Cell Tech Therapeutics Ltd (Slough, UK). Anti-phospho-SLP-76, pTyr128 and anti-VE-cadherin were purchased from Cell Signaling Technologies (Danvers, MA, USA). Anti-HIF-1, anti-pZAP-70-FITC, anti-pZAP-70-PE, anti-LAG3, anti-TIM3, anti-PD-1, eFluor 670, anti-CD45-APC, anti-CHS-PerCP, anti-CCR7 and anti-CD45RO were purchased from BD Biosciences (Franklin Lakes, NJ, USA). Anti-talin and anti-vinculin were purchased from Santa Cruz Biotechnology (Dallas, TX, USA). Goat anti-human Fc-PE was purchased from Millipore (Billerica, MA, USA). Alexa Fluor (AF) labelled secondary anti-mouse, anti-goat anti-rabbit antibodies, streptavidin AF 488 antibody, Texas Red Phalloidin and AF488 Phalloidin were purchased from Life Technologies (Carlsbad, CA, USA). Anti-Fab DyLight 488 was purchased from Jackson ImmunoResearch (Suffolk, UK).

Recombinant proteins. Recombinant human TNF α , IFN γ , TGF β and IL-6 were purchased from Genentech Inc. (San Francisco, CA, USA). Human IL-1 α was purchased from Sigma-Aldrich (St. Louis, MO, USA). IL-7 and IL-15 were purchased from Peprotec (Rocky Hill, NJ, USA). ALCAM-Fc and HER2-Fc were purchased from R&D Systems (Minneapolis, MN, USA).

Chemicals. BIRT 377 was purchased from Tocris Cookson (Bristol, UK).

Image analysis of primary glioblastoma and medulloblastoma samples and normal brain tissue. Tissue microarrays and individual tumour frozen sections were scanned on a Leica microscope with a 20 \times objective or a Cytation5 with a 10 \times objective. Images were acquired in the DAPI, ALCAM-FITC, CD31-TRITC and ICAM1-Cy5 or VCAM-Cy5 channels, to create a 37 \times 37 \times 4 image data set. Images (5–20 fields of view) were sorted and batch-processed using a custom-made Matlab script. FITC, TRITC and Cy5 images were background-corrected by top-hat filtering, and signal noise was removed using the adaptive Wiener method. The positive signal in each channel (ALCAM-FITC or ICAM-cy5 or VCAM-cy5) was then segmented by channel-specific intensity thresholding and tissue masking. Regions of interest were gated based on CD31 fluorescence, and each pixel in identified regions was given a fluorescence intensity value. To confirm ALCAM as an endothelial marker, co-localization analysis between CD31-TRITC and ALCAM-FITC channels and also for control CD31-TRITC and ICAM-Cy5 was done by measuring the Manders co-localization coefficients (MCC⁴³), using a MATLAB written algorithm. MCCs were then calculated between FITC signal (G) and TRITC or Cy5 signal (R) as:

$$M_1 = \frac{\sum G_{i, \text{colocal}}}{\sum G_i}$$

where $G_{i, \text{colocal}} = G_i$ if $R_i > 0$ and $G_{i, \text{colocal}} = 0$ if $R_i = 0$; and

$$M_2 = \frac{\sum R_{i, \text{colocal}}}{\sum R_i}$$

where $R_{i, \text{colocal}} = R_i$ if $G_i > 0$ and $R_{i, \text{colocal}} = 0$ if $G_i = 0$.

As a negative control, FITC images were rotated by 180° before calculating the MCCs.

Cell culture. *Primary brain tumour endothelial cells (pTECs).* Primary GBM surgical specimens and other primary tumour samples were obtained on a human protocol approved by the Institutional Review Board (IRB) of Baylor College of Medicine, The Houston Methodist Hospital and Texas Children's and Toronto Children's Hospitals, and were stained or processed to isolate patient-derived pTECs. Patients and healthy donors gave consent before the samples were obtained unless these were from publicly available data sets. In brief, the tissues were minced in DMEM (Gibco, Waltham, MA, USA) and digested with 1 mg/ml collagenase and 10 mg/ml DNase (BD Biosciences, San Diego, CA) in DMEM for 30 min at 37°C with continuous shaking at 180 r.p.m. The tissue suspension was filtered through a mesh screen (200 μ m), washed with Dulbecco's phosphate-buffered saline (D-PBS), centrifuged at 1,000g for 10 min at 4°C and the pellet was re-suspended in DMEM containing 20% BSA. A second digestion was done for 2 h on a shaker at 180 r.p.m. using collagenase and dispase (1 mg/ml) following which the digested layer was separated on 40% percoll gradient and centrifuged at 700g for 10 min. The endothelial cells were taken from the cloudy interphase, washed in D-PBS and grown in brain endothelial cell medium (EBM; Lonza, Basel, Switzerland) supplemented with 20% fetal bovine serum (FBS), bFGF (20 μ g/ml), heparin (100 μ g/ml) and puromycin (500 μ g/ml) on fibronectin (0.4 mg/ml) and collagen (0.1 mg/ml) pre-coated plates.

After 7–10 days, cells were checked for endothelial cell markers CD31, von Willebrand Factor and VE cadherin. They were sorted on the pan-endothelial

marker CD31 at the Baylor College of Medicine Flow-cytometry Core. The viable sorted CD31⁺ propidium iodide (PI)-negative population of isolated GBM endothelial cells was regrown in EBM in fibronectin/collagen-coated flasks and used at passage 1–2 for the experiments.

Other primary and established cell lines. Primary GBM cell lines were isolated from surgically excised GBM specimens as described⁴⁴. Human brain microvascular endothelial cells (HBMECs), HBVPs and CD1⁺ astrocytes were obtained from ScienCell Research Laboratories (Carlsbad, CA, USA). Human lung microvascular endothelial cells (HMVEC-Ls) were obtained from Lonza (Basel, Switzerland). Human umbilical cord vascular endothelial cells (HUVECs), human embryonic kidney-293A (HEK293) cells, U87-GBM cell line mouse cerebral microvascular tumour endothelial cells, bEnd.3, and the mouse lymph node vascular endothelial cells 2-H11 were obtained from ATCC (Manassas, VA, USA). BALB/c mouse primary brain microvascular endothelial cells and BALB/c primary mouse lung microvascular endothelial cells (PMVECs) were obtained from Cell Biologics Inc. (Chicago, IL, USA).

Astrocytes and brain endothelial cells were cultured in EBM supplemented with 20% FBS and growth factors using supplement kit EGM-2 BulletKit containing VEGF, ECGS, heparin, EGF, hydrocortisone, L-glutamine (2 mM) and puromycin (4 mg/ml). HEK293 cells, U87-GBM cell lines and primary GBM cell lines processed in our laboratory were maintained in DMEM supplemented with 10% FBS and 2 mM GlutaMAX-I (Gibco). T cells were maintained in T cell medium (250 ml RPMI-1640, 200 ml CLICKS with 10% FCS containing 2 mmol/l GlutaMAX-I).

All cells were cultured in an incubator at 37°C in a humid atmosphere of 5% CO₂ and passaged at 70% confluency. All endothelial cells were routinely analysed for CD31 and vWF expression by flow cytometry before use.

Immunohistochemistry and immunofluorescence. Expression of ALCAM on primary GBM and medulloblastoma tissue sections was detected by immunohistochemistry on formalin-fixed paraffin-embedded (FFPE) slides. Following deparaffinization and rehydration, endogenous peroxidase activity was blocked with 30% H₂O₂ and antigen retrieval was performed using Dako antigen retrieval solution (BioGenex, Fremont, CA, USA) for 90 min at 90°C under pressure. Avidin, biotin (BioGenex) and Fc receptor (Innovex Biosciences, Richmond, CA, USA) blocking reagents were applied to the sections before a 4°C overnight incubation with rabbit anti-human anti-ALCAM (1:100). The sections were developed with HRP conjugated anti-rabbit antibody (1:1,000; Abcam) using DAB as chromogen (BioGenex). All slides were counterstained in Harris haematoxylin, dehydrated and mounted. Images were acquired using an Olympus light microscope. Scoring of CD3 positive DAB signal was analysed using IHC_Profiler plugin in ImageJ. Images were scored by a pathologist blinded to the conditions.

Frozen tissue sections were also made from primary GBM surgical specimens at the Baylor College of Medicine pathology core and used for co-staining of CD31 and ALCAM. The slides were fixed with cold acetone/methanol for 15 min, antigen retrieval performed using 1 \times citrate buffer (Thermo Fisher Scientific), blocked with human goat serum for an hour and probed with mouse anti-human CHS1 (1:100) and rabbit anti-human ALCAM antibody (1:50) overnight. Slides were then incubated for an hour with secondary goat anti-rabbit Alexa-fluor 488 and goat anti-mouse Alexa Fluor 647, respectively, counterstained with DAPI and mounted. Tissue images were taken using a Zeiss confocal spinning disk microscope (Zeiss, Oberkochen, Germany) at 40 \times magnification.

Flow cytometry for surface expression of brain endothelial CAMs. For expression of CAMs, 1 \times 10⁶ HBMECs and pTECs were harvested and stained using anti-human ALCAM-PE, anti-human ICAM1-FITC and anti-human VCAM1-FITC. Expression was analysed by flow cytometry at basal levels, after conditioning with IL-6 (100 ng/ml) TNF α (10, 100, 500 ng/ml), TGF β (1 μ g/ml), and GBM supernatant for 6 h. FlowJo data analysis software (FLOWJO, LLC, Ashland) was used for all flow cytometric analyses.

ALCAM expression during basal and pathological conditions was measured on mouse (2-H11, the BALB-C primary brain endothelium bEnd.3 and PMVECs) and human endothelial cells (HBMECs, pTECs, HMVEC-Ls, HUVECs) by flow cytometry. For inflammatory conditions, cells were cultured with TNF α , TGF β , IL-6, IL-1, IFN γ at optimized concentration of 100 ng/ml for 6 h. To simulate tumour environment, endothelial cells were exposed for 6 h to fresh supernatants from GBM cell culture, and collected 24 and 48 h after addition.

Conditioned and normal cells (1 \times 10⁶) were washed in PBS containing 2% FBS and 0.1% sodium azide (FACS buffer; Sigma Aldrich) and stained with ALCAM-PE for an hour in the dark along with matched isotype controls. Approximately 100,000 events per tube were captured using a GalliosTM flow cytometer (Beckman Coulter Inc., Brea, CA) or BD AccuriTM C6 (Becton Dickinson, Franklin Lakes, NJ) and data analysed by Kaluza software (Beckman Coulter Inc.) or FlowJo data analysis software (FlowJo LLC, Ashland, OR).

Western blotting. Conditioned and non-conditioned endothelial cells (1 \times 10⁶) were lysed with RIPA lysis and extraction buffer (ThermoFisher Scientific) per

manufacturer's recommendations and 10 µg each of protein extract was separated by SDS-PAGE, blotted onto PVDF membranes (GE Healthcare, Buckinghamshire, UK) and probed with primary antibodies against ALCAM (1:1,000) and β -actin (1:1,000) at 4 °C overnight. Following incubation with HRP-conjugated secondary anti-mouse (1:25,000) and anti-rabbit antibody (1:5,000), respectively, the blots were developed with ECL prime western blotting detection reagents (GE Healthcare, Chicago, IL, USA). Analysis was done using Image J (NIH) and ALCAM expression was normalized to β -actin.

HS design, synthesis and production of HS T cells. Minimal binding to ALCAM was mapped in silico to domain 3 (D3) of native CD6 and the adjacent stalk (ST). Using Clone Manager (Sci-Ed Software, Cary, NC), the HS prototype molecule was designed as a leader sequence followed by an exodomain formed of domain 3 plus the stalk, followed by an IgG1 hinge, connected to CD6 transmembrane and a CD6 endodomain, formed of the full-length CD6 signalling domain (amino acids 30–400, NC_000011.10 (60971641..61020377)). Subsequently, multimers of the prototype (a trimer (3HS) and a pentamer (5HS)) were generated using multiples of the exodomain. Furthermore, to study the signalling domain function, truncated versions (HSA, 3HSA and 5HSA) were designed with a stop codon placed after 21 amino acids proximal to the transmembrane domain. Expression-optimized DNA sequences with exodomain wobbled in multimers were synthesized by GeneArt Inc. using oligonucleotides, and cloned into the Gateway entry vector pDONR221. Then each HS construct transgene was cloned in the correct frame into an SFG retroviral vector and sequences were verified. For the in vivo bioluminescence tracking, all HS sequences were followed by a 2A sequence and a GFP and firefly luciferase fusion transgene.

To produce retroviral supernatant, 293T cells were co-transfected with SFG retroviral vector, Pef-Pam-e plasmid encoding the sequence for MoMLV gag-pol, and plasmid pME-VSVG containing the sequence for VSV-G, using GeneJuice transfection reagent (EMD Biosciences, San Diego, CA). Retroviral supernatants were collected 48 and 72 h later and cryopreserved. OKT3/CD28-activated T cells were transduced with retroviral vectors as described⁴⁴. In brief, peripheral blood mononuclear cells (PBMCs) were isolated by Lymphoprep (Greiner Bio-One, Monroe, NC) Ficoll gradient centrifugation. 5×10^5 PBMCs per well in a 24-well plate were activated with OKT3 (OrthoBiotech, Raritan, NJ) and CD28 monoclonal antibodies (BD Biosciences, Palo Alto, CA) at a final concentration of 1 µg/ml.

For transduction, a non-tissue culture treated 24-well plate was pre-coated with a recombinant fibronectin fragment (Retronectin; Takara Bio USA, Madison, WI). Wells were washed with PBS and incubated twice for 30 min with the retrovirus supernatant. Subsequently, 2×10^5 T cells per well were transduced with retrovirus in the presence of IL-7 at 10 ng/ml and IL-15 at 5 ng/ml. After 48–72 h cells were removed and expanded in G-rex (Wilson Wolf, St Paul, MN) in T cell medium supplemented with IL-7 at 10 ng/ml and IL-15 at 5 ng/ml for 10–15 days before use.

Transduction efficiencies were assessed with flow cytometry using mouse anti-human D3 monoclonal antibody followed by goat anti-mouse Alexa-fluor 488 conjugate, or human ALCAM-Fc followed by a APC-conjugated goat anti-human Fc. Secondary only controls were incorporated and for all transductions percentage was normalized in comparison to native expression of CD6 on non-transduced (NT) control T cells.

HER2-CAR HS T cells used in the anti-tumour in vivo experiment were generated by co-transduction with FRP5 HER2-specific scFv^{34,44} and the 5HS construct. Surface expression was detected with flow cytometry using goat anti-mouse Fab fragment specific antibody conjugated with DyLight 488 and HS mAb followed by goat anti-mouse AF 488.

Reverse transcriptase polymerase chain reaction (RT-PCR). Using RNeasy extraction Kit (Qiagen), total RNA was extracted from 1×10^6 T cells 8 days after electroporation with control Cas9 only or with gRNA. 4 mg of pre-treated RNA with an RNase-free DNase and was used for cDNA synthesis by using the SuperScript III First-Strand Synthesis System (Life Technologies). Aliquots of the RT product were used for regular RT-PCR amplification for ALCAM and GAPDH as positive control. The reaction was carried out in a total volume of 20 µl containing 3 µl reverse-transcribed cDNA, 1 unit of Q5 high fidelity Taq polymerase (biolab), and 20 µM of each primer. For ALCAM cDNA, the forward primer, 5'-GTCTGGGCAATAGTGACT CC-3' and reverse primer 5'-AACCATGCAAGTGGAAA CC-3 were used. For the control housekeeping gene, GAPDH, the forward primer 5'-TGACCACCACTGCTTAGC-3' and reverse primer 5'-GGCATGGACTGTGGTCATGAG-3' were used. The resulting gene amplicons were analysed by agarose gel electrophoresis; ALCAM at 490 bp and GAPDH at 660 bp.

Flow cytometry. T cell phenotype, exhaustion and proliferation. For phenotype analysis of the transminating T cells, 2×10^5 T cells, all HS T cell groups and the control T cells were stained with anti-human CCR7-PE-Cy7, and anti-human CD45RO-PE for 60 min at room temperature. Similarly, for exhaustion analysis, 2×10^5 T cells were collected from the bottom chamber after transmigration and washed with PBS then incubated with LAG3, TIM3 and PD-1 antibodies.

Versacomp antibody capture beads (Beckman Coulter, Brea, CA) were stained with the same antibodies to allow accurate compensation. For proliferation analysis, 2×10^5 T cells were collected before and after transmigration, and cells were washed and labelled with eFluor 670 diluted to 10 µM and incubated for 25 min at 37 °C in a water bath in the dark before the assay, according to the manufacturer's guidelines. Proliferation analysis and proliferation index comparison were done using FlowJo software.

T cell signalling. For pZap70 and talin detection, 2.5×10^5 T cells, all HS T cells versus control T cells, were fixed and permeabilized using 4% PFA and 0.02% tween-20 then stained with anti-human pZap70-FITC and mouse monoclonal talin followed by anti-rabbit Alexa fluor 488. For LFA-1 open confirmation surface staining, 2.5×10^5 T cells were collected before and after transmigration and stained for LFA-1 specific extended confirmation using biotin labelled KIM127 (recognizes the extended integrin β -chain) followed by a secondary streptavidin-FITC antibody. More than 100,000 events were acquired using Accuri C6 (Becton Dickinson, Franklin Lakes, NJ). FlowJo data analysis software (FLOWJO, LLC, Ashland) was used for all flow cytometric analyses.

Confocal microscopy. SLP-76. 50,000 GFP-tagged HS cells and Δ HS cells were collected and then seeded for 2 h over a glass chamber slide (Thermo Fisher Scientific) pre-coated overnight with 1 µg ALCAM, then fixed and permeabilized with Fixperm solution quenched with ammonium chloride, then incubated in blocking solution (PBS containing 0.01% Triton X-100 with 1% BSA). Finally, cells were immunolabelled with anti-phospho-SLP-76 (pTyr128) followed by anti-mouse Alexa 647. Clustering images of SLP76 at the HS-ALCAM interface were captured using a Zeiss Axio-Observer Z1 confocal microscope equipped with a Yokogawa CSU10 spinning disc, a Zeiss 63 \times /1.43 NA objective, and a Hamamatsu Orca-AG camera. Single 0.1 µm z-slices with SLP-76 puncta at the eGFP-G-expressing T cell surface were detected and compared to Δ HS cells.

Talin-LFA-1-ICAM1 colocalization experiments. The endothelium-HS T cell interface was imaged using confocal microscopy. HS cells and control T cells were seeded over an HBMEC monolayer, pre-stimulated with TGF β (0.1 µg/ml) for 4 h to ensure ALCAM expression, and incubated for 2 h at 37 °C to allow conjugate interactions. Conjugates between HS cells and HBMECs were then fixed and permeabilized with 4% paraformaldehyde and 0.02% saponin (Tween), blocked with PBS containing 0.01% Triton X-100 with 1% BSA and stained for LFA-1 open conformation, ICAM1, and talin. Primary antibodies were subsequently detected by anti-mouse AF 647, AF 488 secondary antibodies and anti-rabbit Alexa fluor pacific blue, respectively. An extra blocking step was performed between the two anti-mouse human antibody staining steps to eliminate background staining. HS cell-HBMEC conjugates were imaged as 0.2-µm Z-steps to cover the entire volume of the podosynapse, determined individually for each conjugate using a Zeiss Axio-Observer Z1 confocal microscope equipped with a Yokogawa CSU10 spinning disc, a Zeiss 63 \times /1.43 NA objective, and a Hamamatsu Orca-AG camera. Images were analysed with Velocity software (PerkinElmer, Waltham, MA). Cluster density of LFA-1, talin, ICAM1 and actin at the interface were calculated using the formula (volume \times MFI) for an equal number of $1 \times 1 \times 1$ -µm voxels selected to cover the interface.

Elisa of TGF β and IL-6 in GBM supernatant. IL-6 and TGF β were quantified using Elisa kits (ab46027) and (ab100647) according to the manufacturer's instructions (on three different samples of GBM supernatant collected 48 h after culture in DMEM without serum).

Cytotoxicity assay. Cytotoxicity assays were performed as previously described²⁰. In brief, to test safety against ALCAM-positive endothelium, U87-GBM and HBMECs were used as targets, and to test fratricide effect, THP1, autologous NT T cells and autologous PBMCs were used. In all experiments: 1×10^6 target T cells labelled with 0.1 mCi (3.7 MBq) ⁵¹Cr at effector to target (E:T) ratios of 40:1, 20:1, 10:1 and 5:1. T cells incubated in complete medium alone or in 1% Triton X-100 were used to determine spontaneous and maximum ⁵¹Cr release, respectively. After 4 h, cells were centrifuged; supernatants collected and radioactivity measured in a gamma counter (Cobra Quantum; PerkinElmer; Wellesley, MA). The mean percentage of specific lysis of triplicate wells was calculated according to the following formula: [test release – spontaneous release]/[maximal release – spontaneous release] \times 100.

Orthotopic GBM and MB xenogeneic SCID mouse model. Tumour establishment. All animal experiments were conducted on a protocol approved by the Baylor College of Medicine Institutional Animal Care and Use Committee and all experiments complied with all relevant ethical regulations. Recipient ICR-SCID mice (C.B-Igh-1^b/IcrTac-Prkdc^{scid}) were purchased from Taconic (Hudson, NY, USA). Male and female 9- to 11-week-old mice were anaesthetized with isoflurane (Abbott Laboratories, UK) followed by an intraperitoneal injection of 225–240 mg/kg xylazine solution and then maintained on isoflurane by inhalation throughout the procedure. After removing hair from the head region, the mice were immobilized in a Cunningham Mouse/Neonatal Rat Adaptor (Stoelting, Wood Dale, IL, USA) stereotactic apparatus fitted into an E15600 Laboratory Standard Stereotactic

Instrument (Stoelting, Wood Dale, IL, USA), and surgery area scrubbed with 1% povidone-iodine. A 10-mm skin incision was made along the midline. The tip of a 31G ½-inch needle mounted on a Hamilton syringe (Hamilton, Reno, NV, USA) served as the reference point. A 1-mm burr-hole was drilled into the skull^{34,44}.

HS T cell homing experiments. For the homing experiment, 5×10^4 unlabelled U87-GBM or DAOY-MB cells in 2.5 µl were injected orthotopically over 5 min. After 10 days of engraftment (based on previous experience xenograft will be vascularized^{45,46}), 10^7 T cells tagged with a GFP Firefly Luciferin fusion gene (eGFP-FFLuc) were injected intravenously through the tail vein. Groups of 15 mice received HS T cells and 10 mice received control NT T cells.

Homing of T cells to brain tumours was assessed by bioluminescence tracking in the brains of the animals using the IVIS system (Perkin Elmer, Akron, OH, USA) after intraperitoneal injection of 300 mg D-luciferin (Perkin Elmer). Mice under isoflurane anaesthesia were imaged individually at the highest sensitivity (level A) for 5 min each at hours 6, 24 and 48, and on days 3, 6 and 8 after T cell administration. Photon emission was quantified using the Living Image software (Perkin Elmer, Akron, OH, USA). A pseudo-colour image representing light intensity (blue: least intense, red: most intense) was generated and superimposed over the grayscale reference image. The regions of signal in the brain were obtained and compared between all test animals ($n = 15$ per group).

In order to assess specific homing and T cell infiltration to GBM xenografts, in a separate experiment, mice ($n = 5$) were euthanized after 24 h and the frontal right lobe containing the tumour was minced and TILs were enriched using a Percoll gradient. Then cells were stained using CD45-PerCp, CHS-APC and mouse anti-human HS mAb followed by anti-mouse AF 647 and analysed using flow cytometry gating concentrically on CD45, CHS and HS positivity. The percentage of TILs was compared between various HS T cell groups and control T cells.

To evaluate the specificity of tumour area homing, random infiltration was evaluated after mincing the left lobe distal from the tumour explants; TILs were isolated and assessed following the method described earlier.

CAR HS T cell efficacy experiments. 5×10^4 eGFP-FFLuc-labelled U87-GBM cells in 2.5 µl were injected orthotopically over 5 min. Three groups of mice were randomized to receive HER2-CAR HS T cells (10 mice per group), HER2-CAR T cells (10 mice per group), NT T cells (5 mice per group) and tumour-only (3 mice per group) at day 6 and 11 days of engraftment through the tail vein. To evaluate the antitumour activity of T cells, tumour sizes were monitored by BLI. Mice were imaged twice weekly for 1 min using the IVIS system (IVIS, Xenogen Corp., Alameda, CA) under isoflurane anaesthesia and 100 µg D-luciferin (Xenogen, Alameda, CA) was injected intraperitoneally. Images were acquired and quantified as described earlier. In order to assess the homing of HER2-CAR HS cells, three

mice that received HER2-CAR HS T cells or HER2-CAR T cells only in a separate experiment were euthanized, lobes containing the tumour area were minced and TILs were assessed as described above.

Mice were regularly examined for neurological deficits, weight loss, signs of stress, and a BLI signal $>10^8$, and euthanized according to pre-set criteria according to the Baylor College of Medicine's Center for Comparative Medicine's guidelines. In none of the experiments were these criteria not fulfilled.

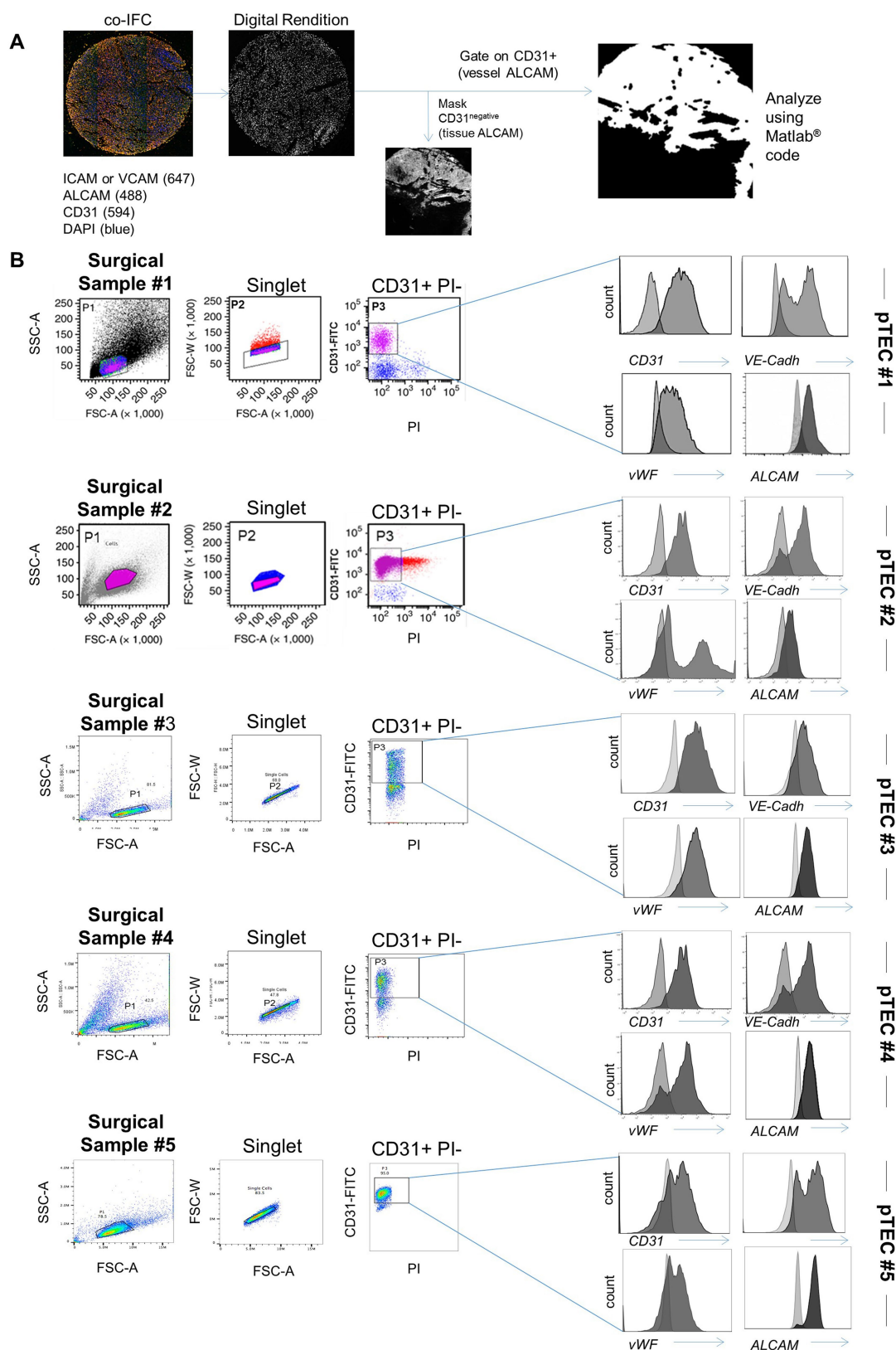
Safety evaluation. The brain, heart, liver, spleen, lungs, stomach, intestine, testicles and kidneys were promptly collected after the mice used in the homing experiments ($n = 5$) were killed and fixed in a 10% formalin solution. Then, the organs were embedded in paraffin, sectioned, and processed for H&E staining and pathologically assessed for histological abnormality or toxicity by a neuropathologist. IHC was carried out on 3 µm brain tumour tissues of different groups ($n = 3$ per group) and probed using rabbit anti-human CD3 (Abcam).

Statistical analysis. Data were summarized using descriptive statistics. Comparisons between groups were carried out using one-way ANOVA or *t*-test. *P* values were adjusted for multiple comparisons using Tukey's test and Dunnett's test when appropriate. The Kaplan–Meier method was used to estimate survival curves and the log-rank test was used to compare the curves. GraphPad Prism 7 software (La Jolla, CA) was used for statistical analysis. The sample size for the animal experiments was calculated on the basis of the primary hypothesis and models derived from pilot studies. Animals were randomized between groups and the operator was blinded to the agents tested. A *P* value of less than 0.05 was considered significant.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

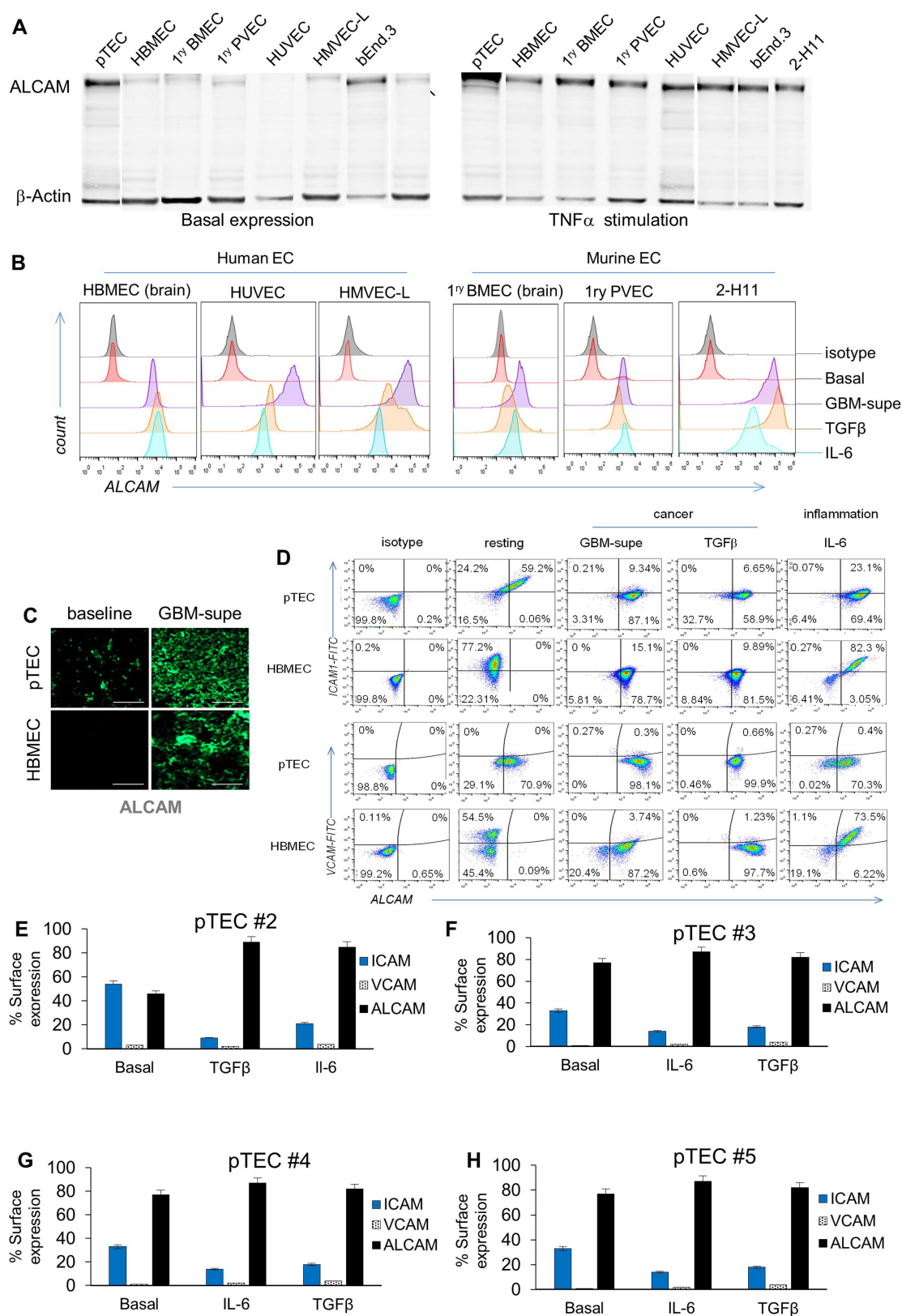
Data availability. All relevant data are included in the manuscript linked as source data; more details are available from the corresponding author on reasonable request

43. Manders, E. M. M., Verbeek, F. J. & Aten, J. A. Measurement of co-localization of objects in dual-colour confocal images. *J. Microsc.* **169**, 375–382 (1993).
44. Ahmed, N. et al. HER2-specific T cells target primary glioblastoma stem cells and induce regression of autologous experimental tumors. *Clin. Cancer Res.* **16**, 474–485 (2010).
45. Sudha, T. et al. Nanoparticulate tetrac inhibits growth and vascularity of glioblastoma xenografts. *Horm. Cancer* **8**, 157–165 (2017).
46. Paris, D. et al. Impaired orthotopic glioma growth and vascularization in transgenic mouse models of Alzheimer's disease. *J. Neurosci.* **30**, 11251–11258 (2010).



Extended Data Fig. 1 | Analysis of CAM expression in primary brain tumours. a, High-throughput IFC analysis of the endothelial adhesion molecules ICAM1, VCAM1, and ALCAM in 93 primary GBM, 25 primary medulloblastoma and 5 normal brain samples. MATLAB segmentation and masking analysis algorithm of the co-immunofluorescence (co-IFC) of ICAM1 or VCAM (acquired on 647 channel), CD31 (acquired on 594 channel) and ALCAM (acquired on 488 channel), and DAPI (acquired on blue/cyan channel). **b**, Isolation and characterization of pTECs.

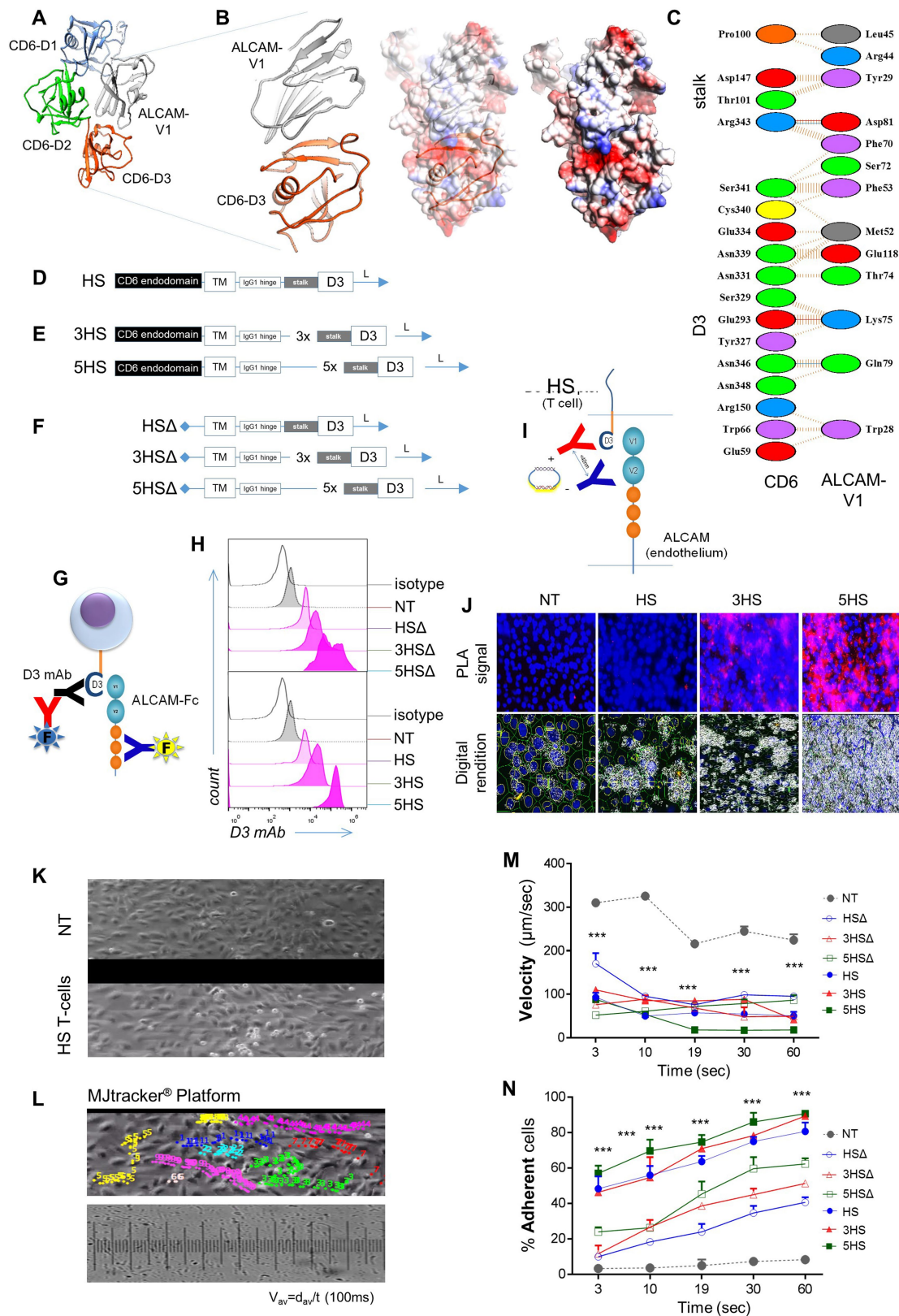
Flow cytometry sorting gating strategy of pTECs from freshly excised glioblastoma (GBM; $n = 5$) based on CD31 positivity. Isolated GBM endothelial cells also expressed the endothelial markers VE-cadherin, von Willebrand Factor (vWF) and ALCAM. Isotype shown in lighter grey and test shown in darker grey in individual histograms. $n = 5$ surgical samples each interrogated at least twice. At least 100,000 events were acquired per condition.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | ALCAM expression in a panel of human and mouse endothelial cells and their reactivity to inflammatory and cancerous conditioning. **a**, Western blot for ALCAM in a panel of human and mouse endothelial cell lines: pTECs, HBMECs, 1ry BMECs, 1ry PVECs, HUVECs, HMVEC-Ls, bEnd.3 (mouse brain tumour EC) and 2-H11 (mouse SV40-transformed axillary lymph node vascular endothelium). Left, basal ALCAM expression except in tumour endothelial cells (pTEC and 2-H11). Right, induction of ALCAMs in all endothelial cells after incubation with $\text{TNF}\alpha$ for 6 h. **b**, Expression of ALCAM at baseline and after 6 h of conditioning in GBM supernatant, $\text{TGF}\beta$ or IL6. Only tumour endothelial cells expressed ALCAM at baseline

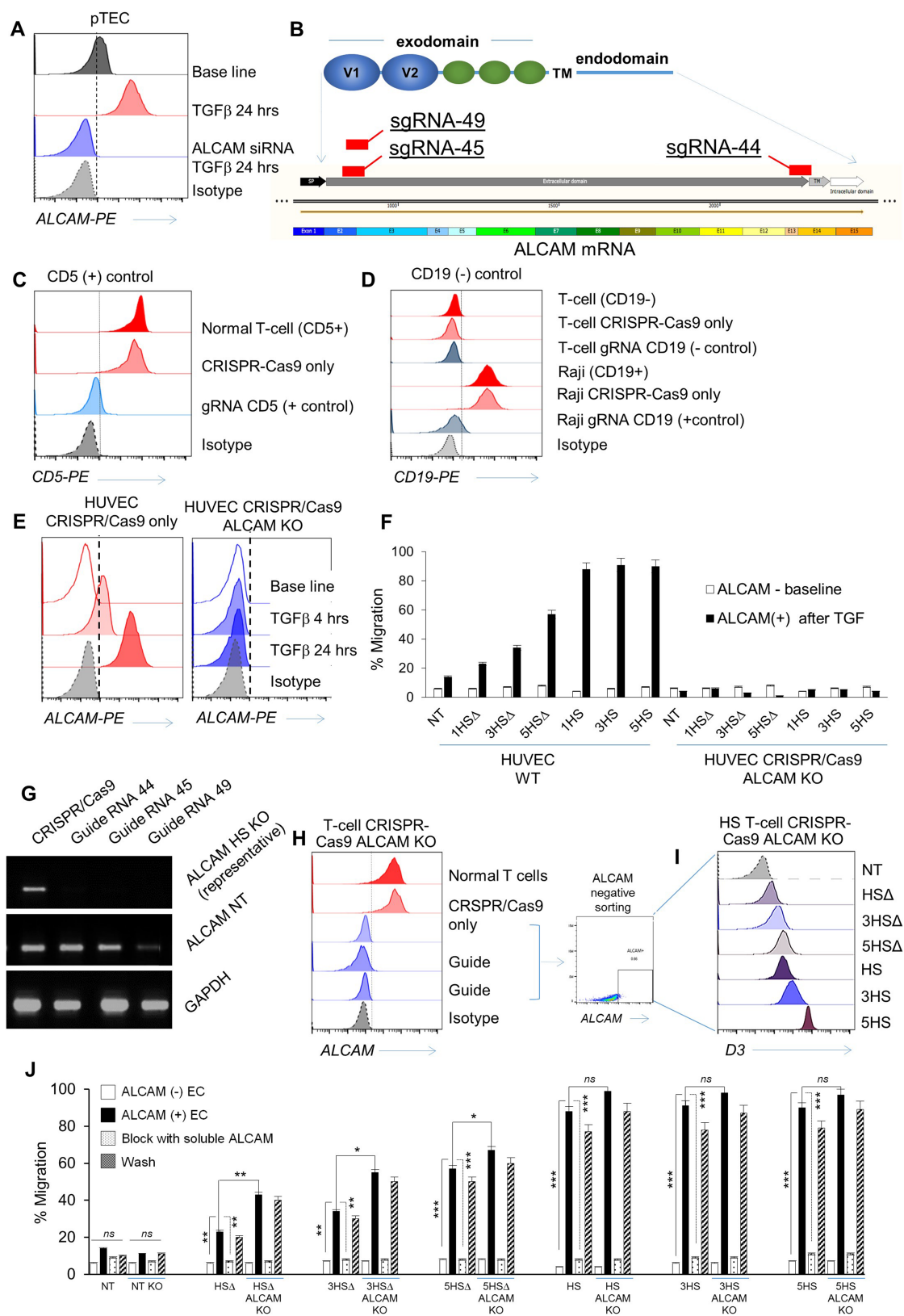
while normal endothelial cells did not. **c**, IFC for ALCAM in 5×10^4 pTECs and HBMECs in the in vitro BBB model at baseline and after culture in GBM supernatant. Scale bars, $50 \mu\text{m}$. **d**, Differential expression of key adhesion molecules at baseline and under the influence of cancer and inflammation in pTECs and HBMECs. Flow cytometry dot plots detailing baseline expression of ALCAM, VCAM1 and ICAM1 on 1×10^4 pTECs and HBMECs and conditioned expression after culture in GBM supernatant, $\text{TGF}\beta$ or IL6. **e–h**, Expression of adhesion molecules at baseline and under the influence of cancer and inflammation in pTECs ($n = 4$) acquired from surgically resected samples (pTEC #1 is shown in Fig. 1g).



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | In silico design of the prototype and derivative HS molecules, their forced expression and detection on T cells, and studies of their in vitro dynamic interactions with endothelial cells under shear stress. **a**, The potential interaction between ALCAM V1 (grey ribbon) and CD6 from computational docking. D1 of CD6 is coloured blue, D2 green and D3 orange. **b**, Details of the potential interaction interface between ALCAM V1 (grey ribbon) and CD6 D3 (orange ribbon). A rendering of the electrostatic surface of ALCAM V1 (grey ribbon) with the D3 domain of CD6 (orange ribbon) in the same orientation. Potential interacting residues are highlighted in the models and in a diagram generated from PDBe PISA and PDBSum (**c**). A small region of positively charged residues in ALCAM V1 appears to interact with a negatively charged patch of residues on CD6 D3. **d**, Structure of the prototype HS molecule. **e**, HS multimers 3HS and 5HS. **f**, HS molecules with non-signalling endodomains, HS Δ , 3HS Δ and 5HS Δ . **g**, Strategy used for surface detection of the HS exodomain using a

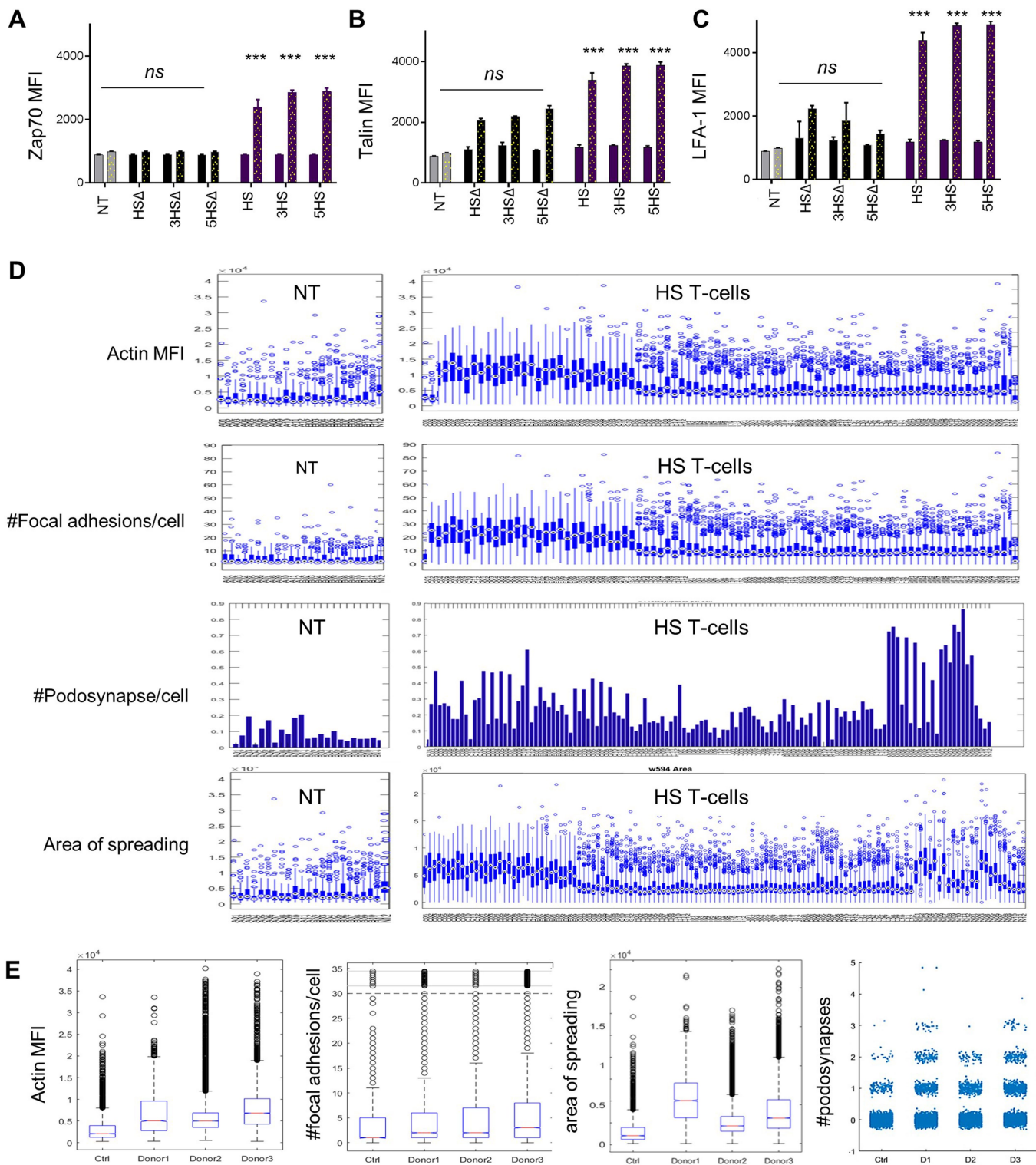
D3-specific antibody and specific binding of the HS exodomain to soluble ALCAM. **h**, Flow cytometry confirming HS surface expression (using D3 monoclonal antibody) on T cells. **i**, Design of the HS–ALCAM PLA experiment. **j**, Digital rendition of PLA using ImageTool. The ALCAM probe (–) binds to the D3 probe (+) to trigger the PCR generating the red fluorescent signal that is quantified as total signal per region (TSR) in Fig. 2f. **k, l**, Dynamic microfluidic studies showing still image from Supplementary Video 1 of Bioflux channels with non-transduced control (NT) T cells (top) versus 1×10^6 HS T cells interrogated under shear force over an ALCAM-expressing endothelium (**k**), and still image from MJtracker demonstrating various T cells under interrogation for various TEM dynamic measures, the standard grid used and the equation used for calculations (**l**). **m**, Dynamic adhesion of T cells to endothelial cells per field of view. **n**, Average dynamic rolling velocity against time; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Two-way ANOVA with Tukey's test for multiple-comparisons (compared to NT cells).



Extended Data Fig. 4 | See next page for caption.

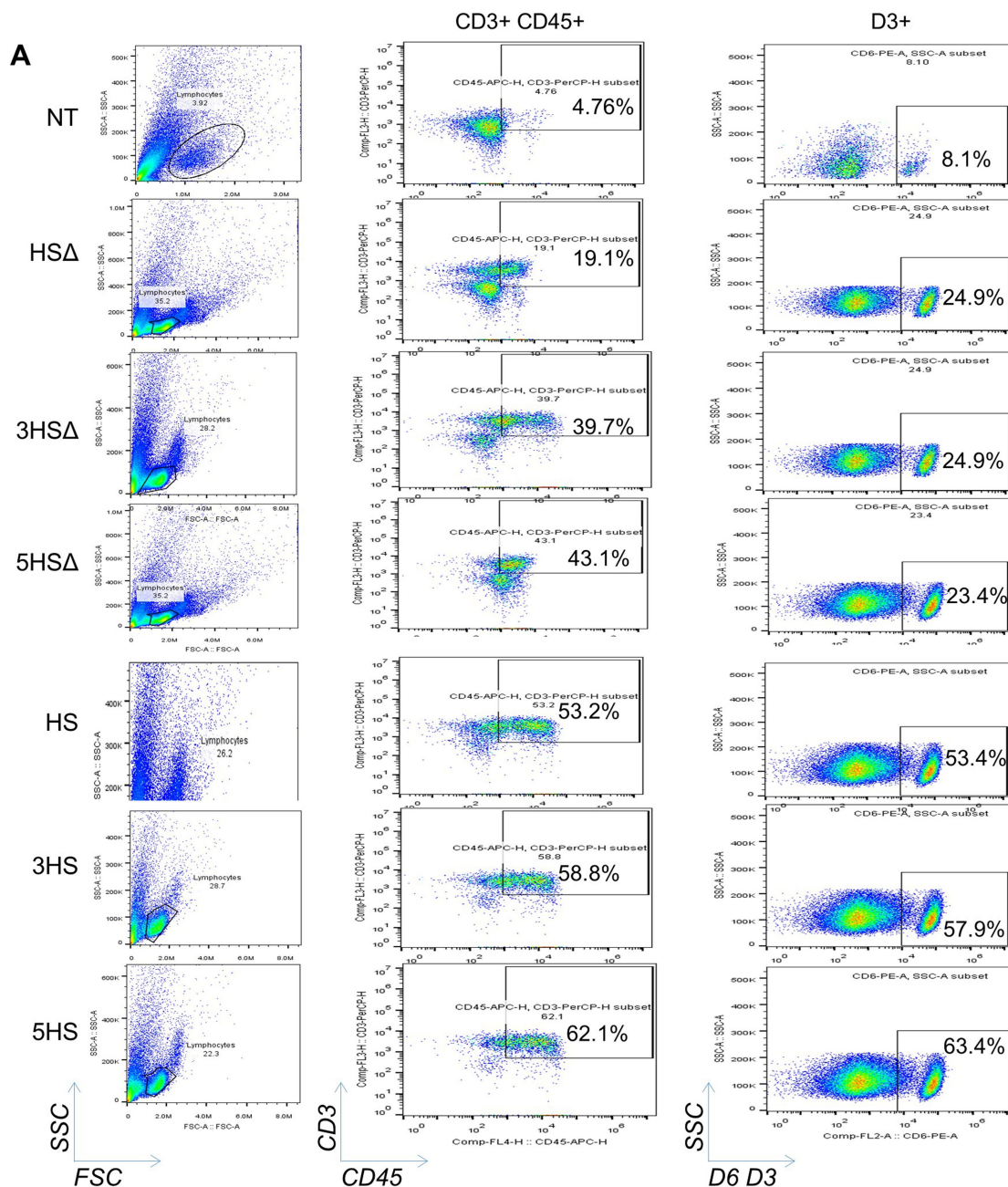
Extended Data Fig. 4 | Functional effects of elimination of ALCAM on endothelial cells and knockout of human ALCAM using CRISPR–Cas9 technology and its effect on T cell migration across the BBB. **a**, Flow cytometry of ALCAM expression on 1×10^6 wild-type pTECs at base-line, after TGF β induction of ALCAM and after being transfected with 25 nM ALCAM siRNA for 48 h to knock down (KD) ALCAM. Transmigration assay using pTECs to simulate a cancerous BBB showing percentage of migrant T cells compared with ALCAM-KD is shown in Fig. 2k. **b**, Highest three scoring guide RNA designs (sgRNA-44, sgRNA-45 and sgRNA-49) as seen on the SnapGene software intended to disrupt ALCAM exons for the extracellular and transmembrane moiety. **c**, **d**, CD5-KO (**c**) and CD19-KO (**d**) were used as positive and negative experimental controls, respectively. **e**, Flow cytometry of ALCAM expression on wild-type HUVECs and HUVEC ALCAM-KO using CRISPR–Cas9 (using the guide sgRNA-45) assessed at baseline and after TGF β incubation. Isotype was used as control. **f**, Transmigration assay showing percentage of 2×10^6 migrating T cells on wild-type HUVECs before and after ALCAM induction

compared to ALCAM-KO HUVECs. Both experiments were done at baseline then after ALCAM induction was confirmed. Data shown as mean \pm s.d. ($n \geq 3$ experiments; donor T cells $n = 3$), $**P < 0.01$, $***P < 0.001$. Tukey's test (compared to wild-type pTECs). **g**, RT–PCR analysis of representative of 1×10^6 ALCAM-KO HS T cells in comparison to wild-type normal T cells. GAPDH was used as an internal control. **h**, Flow cytometry showing $>90\%$ knockout efficiency of the three sgRNAs on 1×10^5 T cells in comparison to wild-type normal T cells; CRISPR–Cas9 only and isotype were used as experimental controls. **i**, Sorted ALCAM-negative KO T cells were then successfully transduced with the six HS constructs. **j**, Transmigration assay showing percentage of 2×10^5 migrating T cells on a cBBB model to compare wild-type with ALCAM-KO T cells in four conditions (ALCAM $^-$, ALCAM $^+$ conditioned with TGF β , after blocking ALCAM, after washing the blocking away). Data in **f** and **j** are shown as mean \pm s.d. ($n \geq 3$ experiments; donors $n = 3$), $*P < 0.05$, $**P < 0.01$. Tukey's test (compared to ALCAM $^+$ T cells).



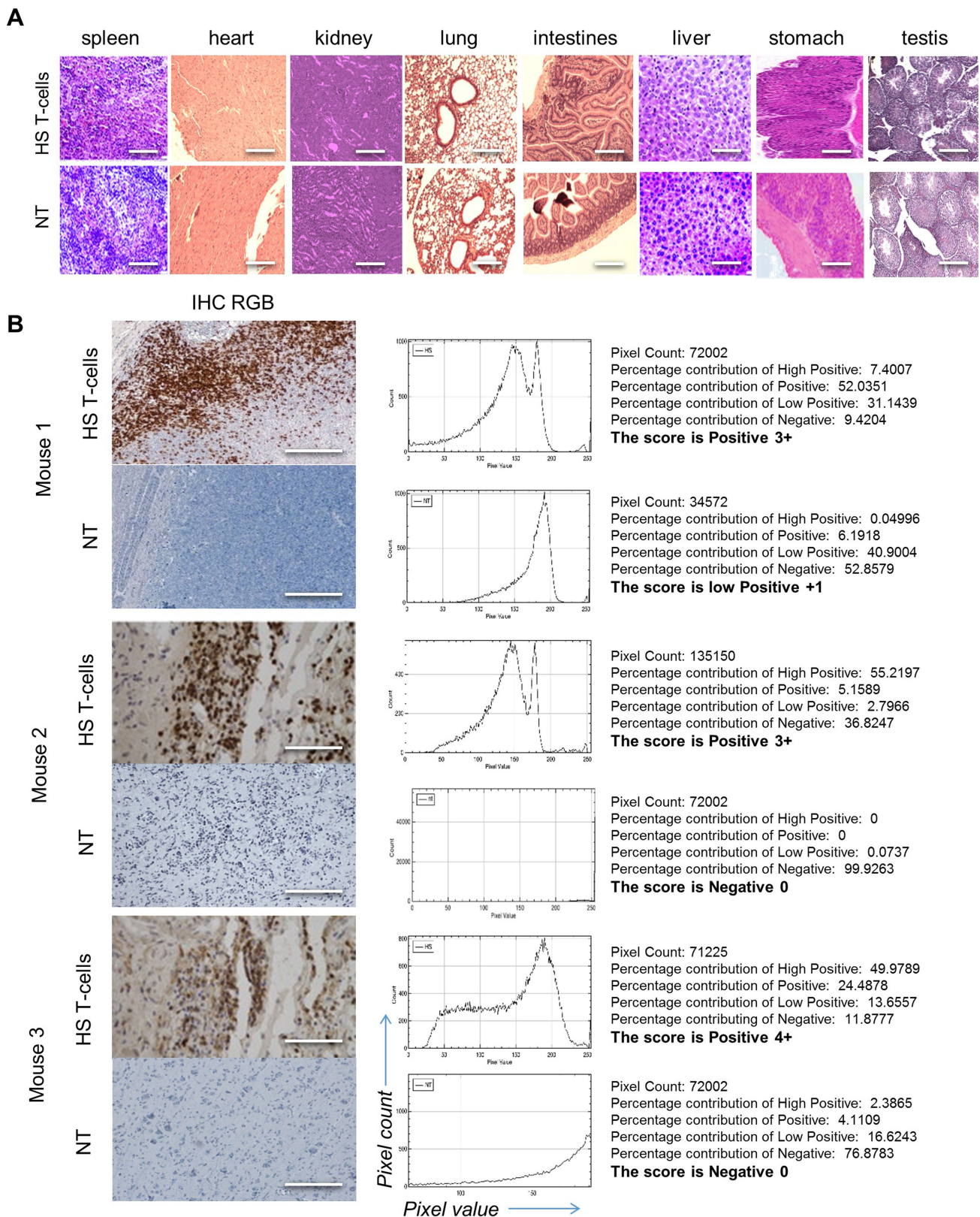
Extended Data Fig. 5 | Flow cytometry quantification of nodes downstream of CD6 signalling endodomains and high-throughput analysis of super-resolution imaging using deconvolution microscopy. **a–c**, Quantification of the flow cytometric data for LFA-1 open configuration (**a**), pZap70 (**b**) and talin (**c**) before (solid bars) and after (dotted bars) TWM of 1×10^5 T cells. *** $P < 0.001$. **d**, Characterization of cellular features of migrant T cells using collective quantification of

actin MFI, focal adhesions at HS–ALCAM interface, area of spreading, and podosynapse formation by high-throughput microscopy in three donors. $n = 200$ –800 cells. **e**, Box plot summary representing single-cell data distributions of all replicates between all three donors expressing HS versus NT controls. Centre lines, data median. Boxes, middle quartiles. Whiskers, upper and lower limits.



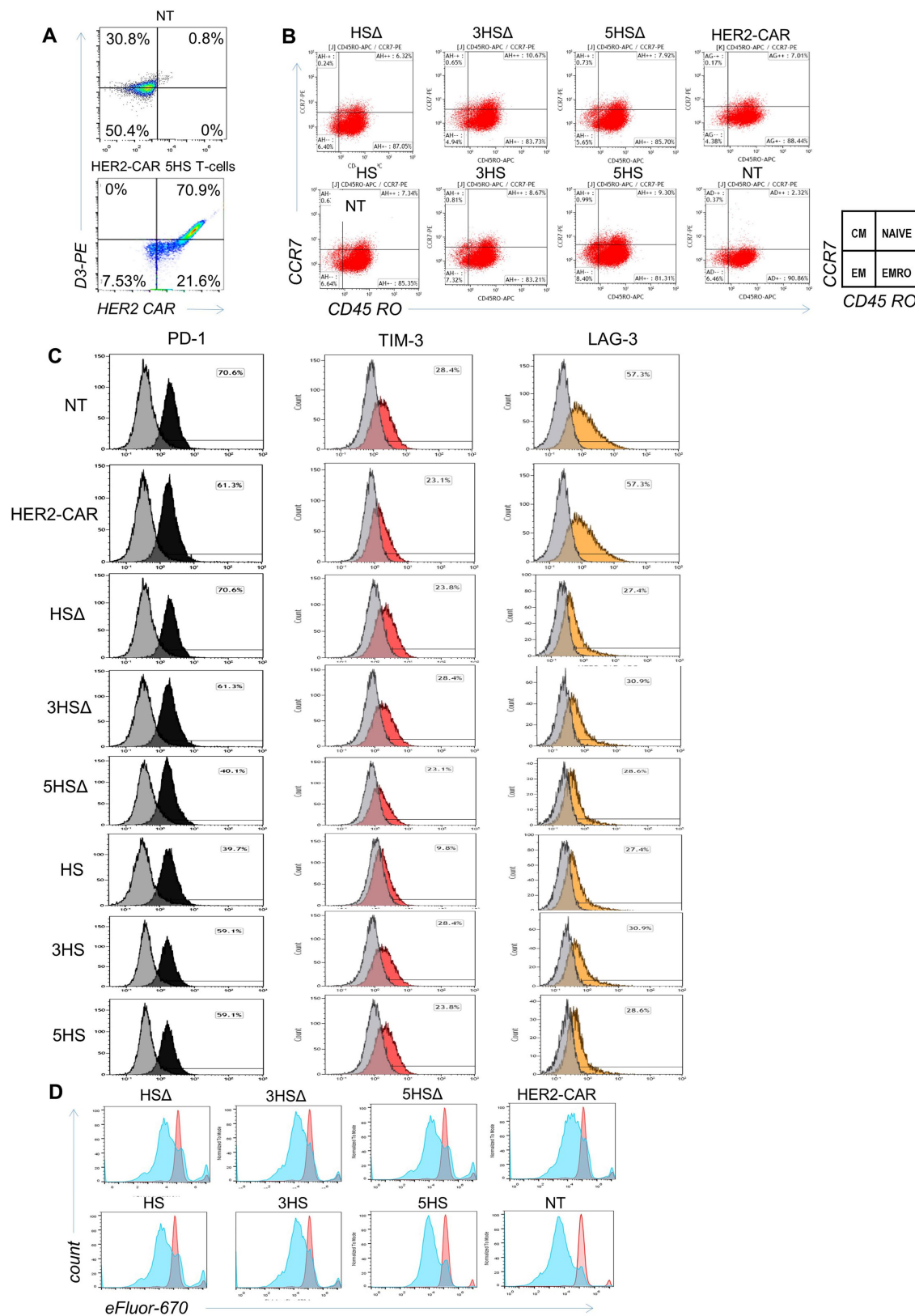
Extended Data Fig. 6 | Assessment of TILs in GBM explants. a, Flow cytometry of 1×10^4 TILs; all HS T cell designs compared with normal T cells gated on CD3⁺CD45⁺ then D3⁺ fractions in GBM explants 24 h

after intravenous infusion. Representative plots shown. $n = 5$ animals per group. **b,** Cranial window on a live mouse bearing U87-GBM tumour (black arrow, right).



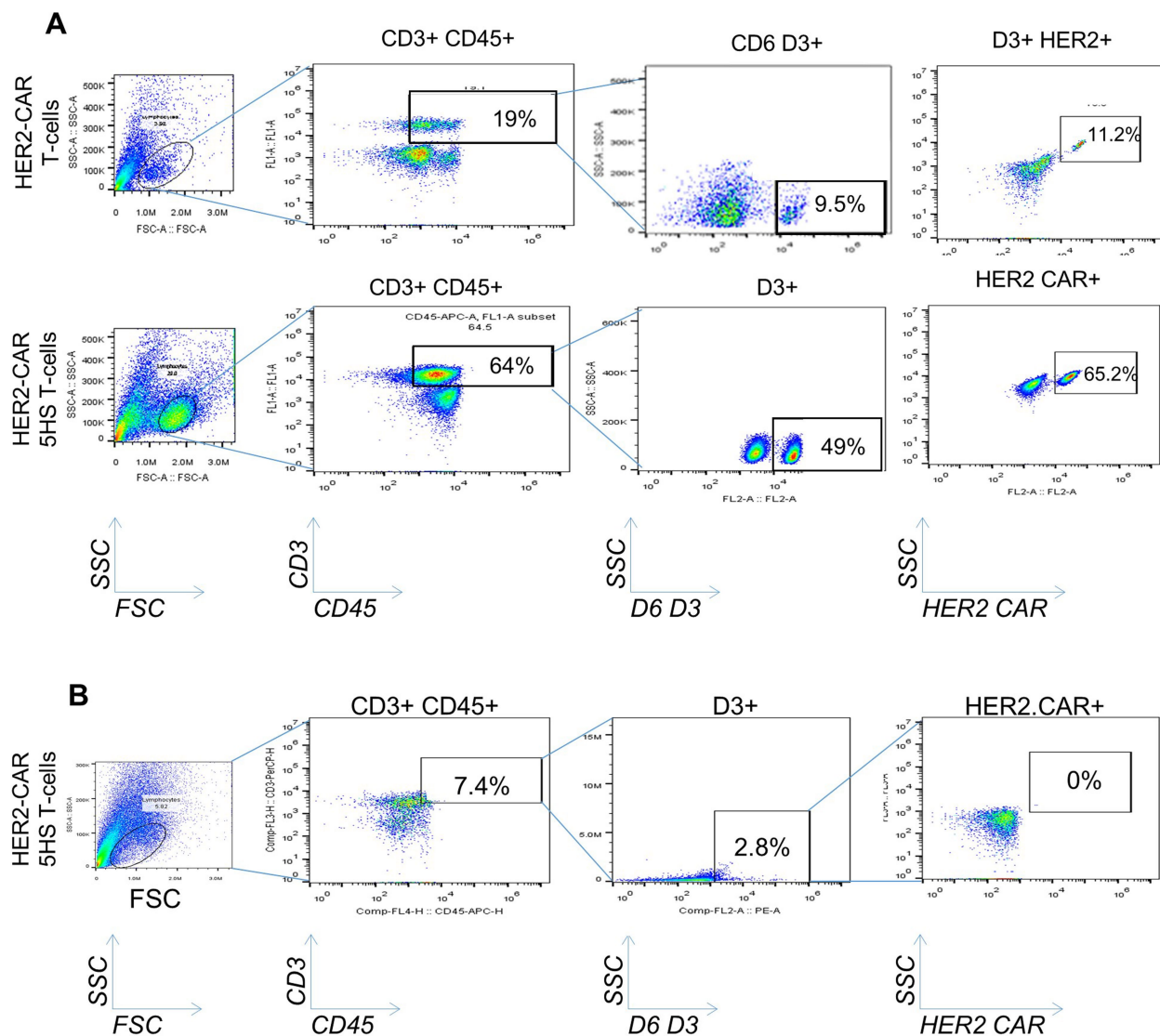
Extended Data Fig. 7 | Analysis of T cell infiltrates in vital organs and normal brain after infusion of HS T cells. a, CD3 immunohistochemistry (IHC) staining of normal vital tissues from animals receiving HS T cells or NT control cells. $n = 3$ mice per group. Scale bars, $40\ \mu\text{m}$. **b,** IHC showing HS T cell infiltrate in micro-dissected GBM xenograft. Scoring of CD3⁺ DAB signal was analysed using IHC-Profiler plugin in ImageJ. Respective

image analysis output and the score assigned using IHC-Profiler are also shown for each image. Total percentage of CD3⁺ DAB signal was more 66% in all mouse brain with HS T cells (scores 3–4) and percentages in control mice were less than 20% (scores 0–1). Scale bars, $50\ \mu\text{m}$. $n = 3$ mice per group.



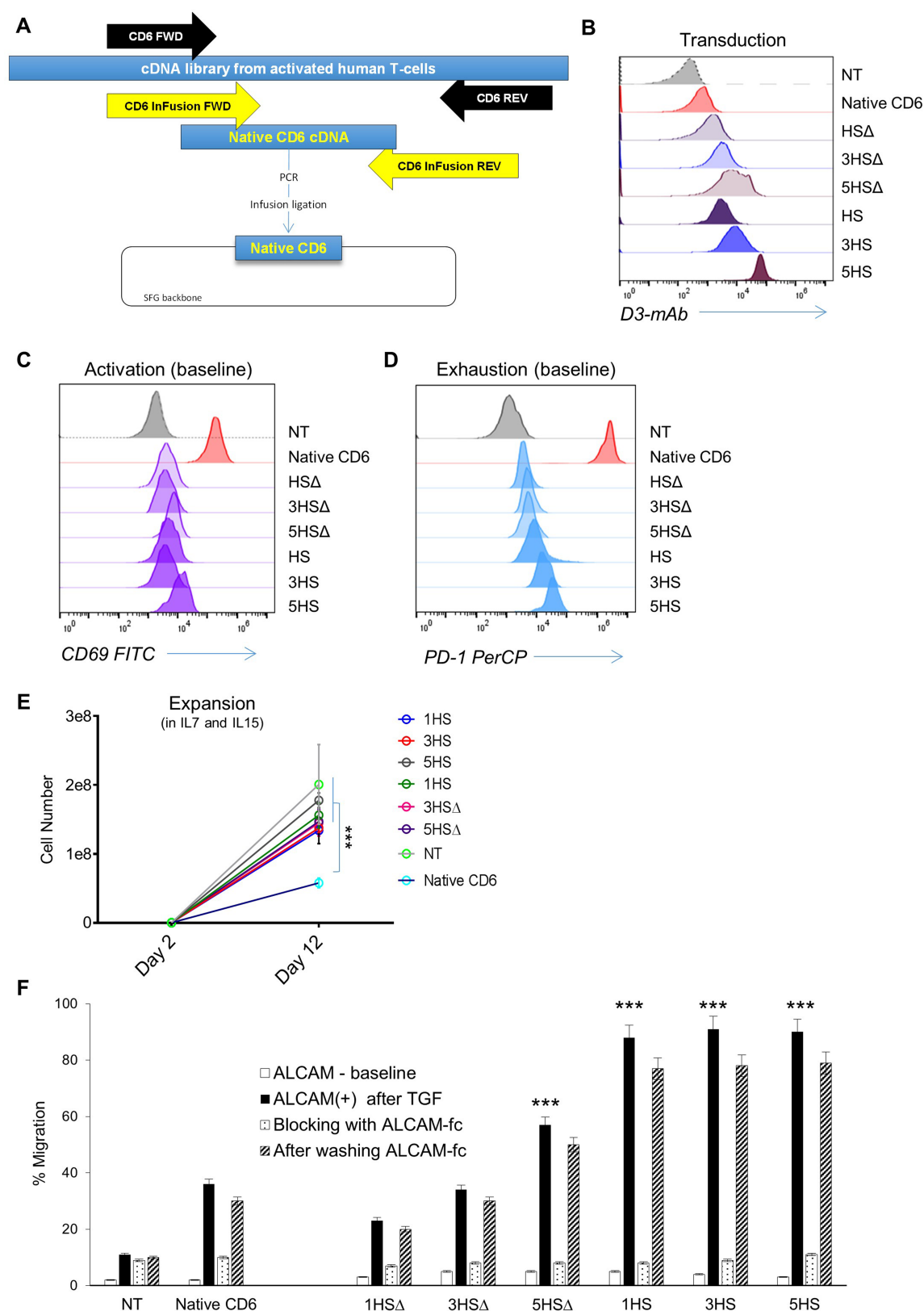
Extended Data Fig. 8 | Characterization of therapeutic T cells after transmigration through an *in vitro* BBB model. a, Flow cytometry assessing HER2-CAR and HS molecule expression in HS HER2-CAR T cells. **b–d**, 1×10^5 T cells were collected from the bottom chamber after transmigration on ALCAM-expressing endothelium and analysed

for CD45RO and CCR7 to assess their centrality (**b**), expression of the exhaustion markers PD-1 (black), TIM-3 (red) and LAG3 (orange) (before transmigration is shown in grey) (**c**), and proliferative capacity before (red) and after (blue) transmigration, using eFluor 670 (**d**).



Extended Data Fig. 9 | Analysis of TILs isolated from tumour xenografts and normal brain for HER2-CAR HS T cells. a, Flow cytometry of TILs isolated from orthotopic tumour xenografts 24 h after intravenous injection of HS T cell products, HER2-CAR T cells and NT control T cells. Xenografts were micro-dissected and TILs were isolated and enriched on a percoll/ficoll gradient. Cells were gated on D3⁺ subset

inside a gate of D3⁺CD45⁺. A subset of HER2-CAR inside a gate of CD3⁺CD45⁺D3⁺ was used to detect HER2-CAR HS T cells specifically. *n* = 5 mice per group, representative data shown. **b,** Flow cytometry following the same gating strategy indicating the absence of HS T cells in the contralateral lobe to the tumour xenograft; data representative of three mice.



Extended Data Fig. 10 | Overexpression of full-length native CD6 and its phenotypic and functional effects on T cells. a, Cloning strategy of native CD6 in an SFG retroviral backbone. **b**, Flow cytometry showing the transduction of 1×10^5 native CD6 relative to HS constructs on T cells. **c**, Flow cytometry of the activation marker CD69 on day 8 after transduction without additional stimulation. **d**, Flow cytometry of the activation and exhaustion marker PD-1 stained with PD-1 PerCP on day 8 transduction at basal level without additional stimulation. **e**, Expansion

plot of T cells expressing the native CD6 relative to NT and various HS T cells; cells were grown in IL-7/IL-15 and collected at day 2 and day 12 post transduction. **f**, Transmigration of 2×10^5 T cells through a cancerous BBB model showing the percentage of migrant T cells expressing native CD6 relative to various HS T cells, and the response to blocking ALCAM and its restitution. Data shown as mean \pm s.d. ($n \geq 3$ experiments; donor T cells, $n = 3$) *** $P < 0.001$ compared to migration of CD6 through ALCAM⁺ BBB. ANOVA with Tukey's post-hoc analysis.

Extensive sex differences at the initiation of genetic recombination

Kevin Brick^{1,4}, Sarah Thibault-Sennett^{2,4}, Fatima Smagulova^{2,3}, Kwan-Wood G. Lam¹, Yongmei Pu², Florencia Pratto¹, R. Daniel Camerini-Otero^{1*} & Galina V. Petukhova^{2*}

Meiotic recombination differs between males and females; however, when and how these differences are established is unknown. Here we identify extensive sex differences at the initiation of recombination by mapping hotspots of meiotic DNA double-strand breaks in male and female mice. Contrary to past findings in humans, few hotspots are used uniquely in either sex. Instead, grossly different recombination landscapes result from up to fifteen-fold differences in hotspot usage between males and females. Indeed, most recombination occurs at sex-biased hotspots. Sex-biased hotspots seem to be partly determined by chromosome structure, and DNA methylation, which is absent in females at the onset of meiosis, has a substantial role. Sex differences are also evident later in meiosis as the rate at which meiotic breaks are repaired as crossovers differs between males and females in distal regions. The suppression of distal crossovers may help to minimize age-related aneuploidy that arises owing to cohesion loss during dictyate arrest in females.

Genetic recombination links homologous chromosomes and facilitates their orderly segregation at the first meiotic division. Recombination is initiated by programmed DNA double-strand breaks (DSBs) that are subsequently repaired as either crossovers or non-crossovers. Recombination frequency and patterning can differ between males and females of the same species: the female crossover rate is higher in humans and mice, and in most studied mammals, crossovers are highly concentrated at sub-telomeric regions in males, but not in females¹. This pattern is not universal, however, and in some species, such as pigs, subtelomeric crossovers are increased in females². Sex differences in recombination have been studied by comparing the genetic end products of recombination, primarily crossovers, between the sexes. However, sex-specific variation at the initiation of meiotic recombination has not been studied. We therefore generated quantitative, high-resolution, genome-wide maps of meiotic DSBs in both male and female mice to examine when and where sex biases in recombination are established, and to determine the mechanism(s) that give rise to these biases.

Sex-specific maps of meiotic DSBs

To map meiotic DSBs in female meiosis, we exploited a method we previously developed³ to map DSB hotspots in mouse^{4–6} and human⁷ males. This variant of chromatin immunoprecipitation followed by sequencing (ChIP-seq) known as single-stranded DNA sequencing (SSDS) detects single-stranded DNA (ssDNA) bound to DMC1 protein, an early intermediate in the DSB repair process^{7,8}. In female mice, meiotic DSBs form in the fetal ovary. Each ovary contains up to 10,000 meiotic cells at the required stage⁹, approximately 100 times fewer such cells than the adult testis; thus, mapping DSBs from a single ovary is not possible. Instead, we mapped DSBs from one pool of 230 fetal ovaries and one pool of 90. From the 230-ovary pool, we generated a DSB map of similar quality to that of 9 independent DSB maps generated from male individuals (Fig. 1a; Extended Data Figs. 1, 2a; signal percentage of tags (SPoT) sample ovary 1 (O1) = 33%; SPoT for testis maps = 22–47%). The DSB map from the 90-ovary pool (sample O2) was of lower quality (SPoT = 6%; Extended Data Fig. 1b) but

shared most hotspots (91%) with the better ovary DSB map (Extended Data Fig. 2).

Most DSB hotspots are found in both sexes (Extended Data Fig. 2a); 88% of hotspots from the better ovary DSB map are found in males, and this increases to 97% of hotspots common to both ovary maps. Hotspots unique to either sex are weak (Extended Data Fig. 2b–d) and contribute less than 2% of the SSDS signal. Given that strength estimates at weak hotspots are noisy and that ChIP-seq provides the relative rather than absolute estimates of hotspot use, it is likely that these hotspots are also used in the other sex, but with a frequency below our detection threshold. Sex-specific hotspots have been described in humans¹⁰; however, this was likely to be an incorrect conclusion from an underpowered study. For example, if we just examine the strongest 5,000 (of more than 13,000) hotspots in each sex, we would erroneously conclude that 30% of hotspots are sex-specific. By contrast, we find few, if any, hotspots that are used exclusively in either sex, in agreement with more recent data from humans¹¹. An intriguing difference between the sexes is that the DMC1 SSDS signal at hotspots appears narrower in females than in males (by approximately 400 base pairs (bp) at the widest point; Fig. 1b, Extended Data Fig. 3). The narrower SSDS signal in females may be explained by shorter DSB end resection, by DMC1 loading over a shorter distance, or by differences in repair dynamics between the sexes. These findings and other evidence^{12–14} indicate that meiotic DSB processing differs between the sexes.

There are notable differences in meiotic DSB repair on the sex chromosomes in males and females. The sex chromosomes share approximately 700 kb of homology (pseudoautosomal region, PAR)¹⁵, and a meiotic crossover at the PAR is required in males, the heterogametic sex. Females have two copies of chromosome X, so crossover formation in the PAR is not essential. Relative to controls, PAR DSBs were enriched in all nine males but in neither of the female samples (Fig. 1c, d). In males, DSBs outside the PAR either remain unrepaired¹⁶ or are continually formed¹⁷ after autosomal DSBs have been repaired. This increases the SSDS signal such that chromosome X hotspots appear stronger in males than in females (Fig. 1e, f). Because of these

¹Genetics and Biochemistry Branch, National Institute of Diabetes, Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA. ²Department of Biochemistry and Molecular Biology, Uniformed Services University of the Health Sciences, Bethesda, MD, USA. ³Present address: IRSET INSERM, U1085, Rennes, France. ⁴These authors contributed equally: Kevin Brick, Sarah Thibault-Sennett. *e-mail: rdcamerini@mail.nih.gov; galina.petukhova@usuhs.edu

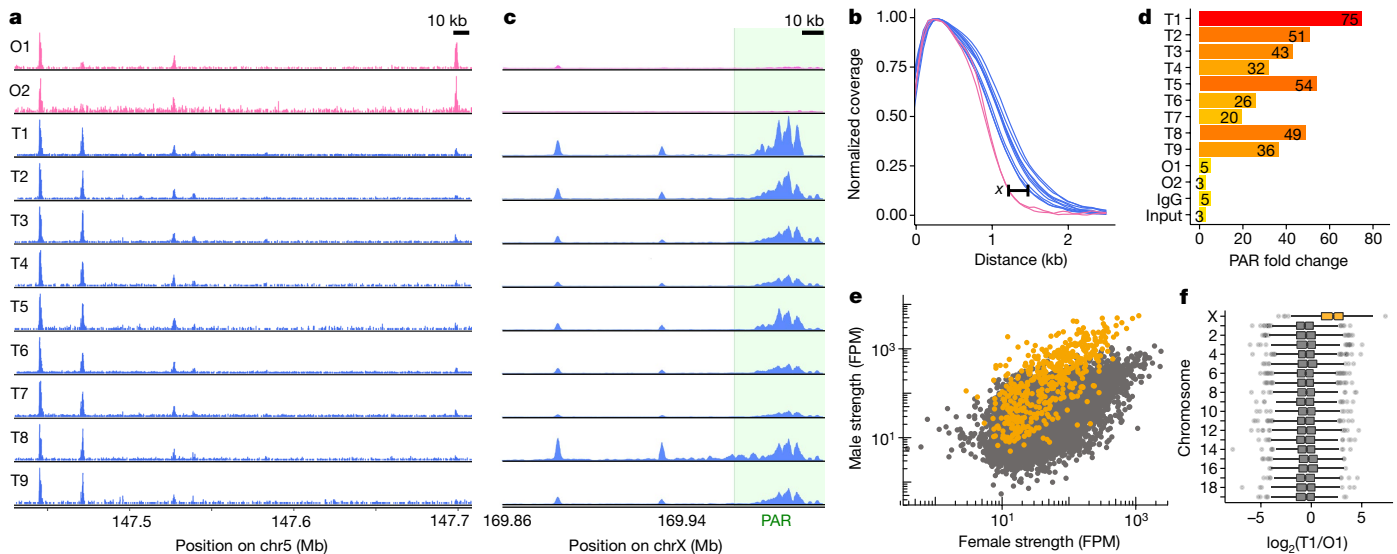


Fig. 1 | Detectable meiotic DSB hotspots in females. a, DSB maps for ovary (pink) and testis (blue). **b**, SSDS coverage (3' distance to centre) at hotspots is narrower in females (maximum difference (x) is approximately 400 bp). **c**, **d**, PAR SSDS signal (green) is enriched in males, but not in females. **e**, **f**, Coverage normalized by total chromosome X (chrX) SSDS.

systematic differences, the sex chromosomes are excluded from subsequent analyses unless explicitly mentioned.

Sex differences at recombination initiation

The SSDS signal at hotspots is highly reproducible in males, with little inter-individual variability (Fig. 2a, c; Spearman's $R^2 \geq 0.90$). Between females, the SSDS signal is also highly correlated (Fig. 2b; Spearman's $R^2 = 0.76$) but to a slightly lesser degree; noise in SSDS estimates for the lower quality O2 map probably reduced this correlation (Extended Data Fig. 2e, f) but stochasticity of hotspot targeting in individual females cannot be ruled out. This is unlikely, however, as these mice are genetically homogeneous and negligible inter-individual variation is seen in males. The SSDS signal at hotspots is markedly different between males and females (Fig. 2c, d; Spearman's $R^2 \leq 0.4$). Indeed, examination of all DSB hotspots found in the best male (testis 1, T1) or best female (O1) sample (20,119 hotspots; see Methods) revealed that 48% of autosomal hotspots are sex-biased ($P < 0.001$, MANORM¹⁸, see Methods; Fig. 2d; $n_{\text{male-bias}} = 4,169$ (22%), $n_{\text{female-bias}} = 5,021$ (26%); $n_{\text{unbiased}} = 9,863$ (52%). The average sex-biased hotspot differed between the sexes by 4.0 ± 4.3 -fold (mean \pm s.d.; median = 2.7-fold), and 1,746 hotspots showed over fivefold difference (Extended Data Fig. 4a). Sex-biased hotspots are probably underdetected, because at stronger hotspots, in which we have the greatest power to detect sex differences, more than 60% of hotspots are sex-biased (Extended Data Fig. 4b). Importantly, sex biases are consistent between the O1 and O2 samples (Fig. 2e, Extended Data Fig. 5a, b), therefore they reflect true sex differences, and not sampling noise in the pooled ovary maps. Approximately $44 \pm 0.4\%$ of the SSDS signal in males and $51 \pm 4\%$ (mean \pm s.d.) in females occurs at hotspots biased to their respective sex (Extended Data Fig. 4c). In addition, a further 16–21% occurs at hotspots biased to the opposite sex (female-biased hotspots in males or male-biased hotspots in females). Therefore, although most hotspots are used in both sexes, most of the SSDS signal, in both sexes, originates at sex-biased hotspots.

SSDS gives an accurate measure of DSB frequency (hotspot strength), tightly correlated with an independent measure of hotspot strength in male mice¹⁶. Nonetheless, because the SSDS signal is probably affected by the lifespan of DSB repair intermediates^{16,19}, a component of the observed sex biases may arise from differential DSB repair dynamics between the sexes. To establish whether sex biases precede DSB formation, we examined histone 3 lysine 4 trimethylation (H3K4me3), a histone modification introduced at hotspots by the PRDM9

zinc-finger protein²⁰. The H3K4me3 signal at hotspots correlated better with the SSDS signal from the respective sex (Extended Data Fig. 5e–j): 69% of female-biased hotspots coincided with an H3K4me3 ChIP–seq peak in fetal ovary, but just 39% of male-biased, and 43% of unbiased hotspots overlapped these sites. Notably, sex-biased hotspots defined using SSDS showed similar sex biases in the H3K4me3 signal (Extended Data Fig. 5f). The magnitude of sex bias is reduced in H3K4me3 ChIP–seq compared to SSDS, perhaps reflecting the reduced sensitivity of H3K4me3 ChIP–seq at hotspots compared to SSDS. Thus, although the contribution of repair dynamics to sex biases remains unclear, sex biases in recombination appear to be established before DSB formation.

PRDM9 defines most DSB hotspots in both sexes; however, the default targeting pathway that targets DSBs to H3K4me3 at functional genomic sites in the absence of PRDM9⁴, is used more frequently by females (Fig. 2f, g; pink and red) than by males (Fig. 2f, g; blue). Increased default targeting may arise because PRDM9 is limiting or because these sites become more accessible in female meiosis. Notably, the recombination (crossover) rate increases locally around functional genomic elements in human females but not males¹¹. However, it remains unclear whether the default targeting pathway is active in humans²¹.

Large-scale influences on sex biases

Chromatin context can modulate DSB hotspot usage^{6,19,22}. Because meiotic chromosome packaging differs between males and females^{23,24}, we looked for evidence of large-scale epigenetic effects that modulate sex biases. Both male-biased and female-biased hotspots occurred in clusters more frequently than expected (see Methods; Fig. 3a–c, Extended Data Fig. 6a–c). Unbiased hotspots did not cluster (Fig. 3a, b). Biased hotspot domains are evenly distributed across chromosomes (Extended Data Fig. 6d) and there seems to be no constraint on cluster size (Extended Data Fig. 6e). A similar proportion of default and PRDM9-defined hotspots were found in clusters (Extended Data Fig. 6f), suggesting that domain-scale regulation of hotspot usage is independent of the mechanism that targets DSBs. Spatial clustering of sex biased crossovers occurs in humans¹⁰, prompting the question of whether these biases occur by a similar mechanism at the initiation of recombination. Deciphering the factors that govern clustering will require genome-wide analyses of meiotic chromosome structure in both sexes.

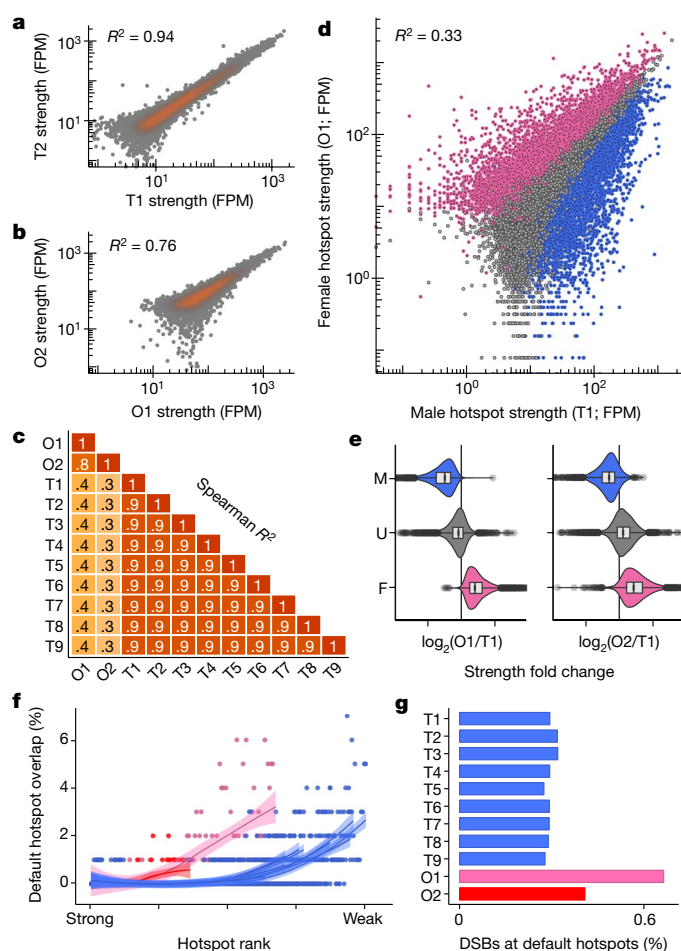


Fig. 2 | Extensive sex differences in DSB hotspot usage. **a–c**, Hotspot strength is consistent between replicates from the same sex (**a**, **b**), but differs between males and females (**c**). **a–c**, Spearman R^2 values are shown. Only hotspots called in both maps are considered. **d**, Strength differs between males and females (female-biased (pink), unbiased (grey), and male-biased (blue) hotspots). **e**, Sex biases are consistent in replicate ovary maps. Curves show LOESS smoothing for each sample. F, female; M, male; U, unbiased. **f**, Default hotspots are used more frequently in females (O1, pink; O2, red; T1–T9, blue). Overlaps were counted in 250-hotspot bins. **g**, More DSBs occur at default hotspots in females.

Recombination in subtelomeric regions is of particular interest because both human and mouse females have decreased distal crossovers relative to males^{11,25}. Distal crossovers in oocytes may be disfavoured as they can increase the risk of chromosome mis-segregation²⁶. The SSDS signal in sub-telomeric regions contrasts starkly with that of crossovers: SSDS is high in females relative to males (Fig. 3d), whereas crossovers²⁵ are less frequent in females (Fig. 3e). Distal crossovers in females may reduce gamete survival²⁶ and be underdetected in pedigree studies. However, in fetal oocytes, in which selection against karyotypic defects should be minimal, crossover depletion was still seen at two subtelomeric hotspots²⁷. Thus, the ratio of crossovers to SSDS decreases close to the telomere in females, but increases in males (Fig. 3f). Coupled with existing data^{27,28}, this suggests that there are fewer crossovers per DSB in distal regions in females. Notably, DMC1 SSDS underestimates DSB frequency in the distal 5 Mb of male mouse chromosomes²⁹. Similar underestimation in females would amplify the crossover-to-DSB deficit in females. Irrespectively, these data indicate that DSB frequency is not the driver of sex differences in distal crossover density.

DNA methylation modulates sex biases

A remarkable difference between the sexes at the time of DSB formation is that the genome is globally demethylated in females but not

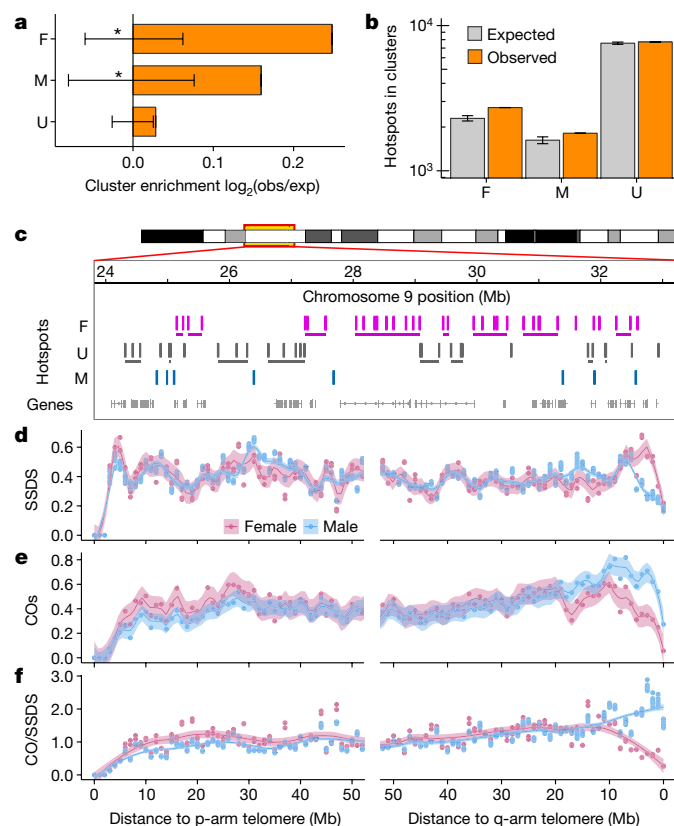


Fig. 3 | Large-scale influences on sex-biased recombination.

a, Sex-biased hotspots cluster more than expected (empirical $*P < 0.001$; see Methods; error bars denote $\pm 99.9\%$ bootstrapped confidence intervals). **b**, Grey denotes median of 10,000 bootstraps; error bars denote 99.9% confidence intervals. Exp, expected; obs, observed. **c**, Clusters of female-biased hotspots (horizontal lines below hotspots). **d**, SSDS decreases in males (blue) relative to females (pink) adjacent to the q-arm telomere. **e**, Crossovers (COs)²⁵ increase in males in this region. **f**, Sex dimorphism in the crossover:SSDS ratio. Profiles in panels **d–f** use 1 Mb non-overlapping windows; the percentage of total signal is shown in panels **d** and **e**. We focus on the q-arm because unassembled centromeric DNA abuts the p-arm telomere in mice.

in males³⁰. DNA methylation can alter the site preferences of DNA-binding proteins³¹; we therefore hypothesized that differential DNA methylation may cause sex biases at the initiation of recombination. Bisulfite sequencing data (Extended Data Fig. 1h, i) revealed that in the testis, the PRDM9-binding site (PrBS) is frequently methylated at male-biased hotspots (Fig. 4a; Extended Data Fig. 7), whereas at female-biased hotspots, DNA methylation instead increases in the region ± 75 bp adjacent to the PrBS. A methylation ‘spike’ at the PrBS 5’ end is common to all hotspots. Neither of the sex-specific patterns is seen at unbiased hotspots, and both patterns are most pronounced at hotspots with greatest sex biases (Extended Data Fig. 8). Meiosis-specific processes do not give rise to these methylation patterns, because qualitatively similar patterns are seen in somatic tissue of both sexes (Extended Data Fig. 8). Importantly, these sites do not escape demethylation in the female germ line (primordial germ cells (oocytes) at 16.5 days post-coitum (d.p.c.)³⁰; see Methods) (Fig. 4a; Extended Data Fig. 8).

The distinct methylation patterns at male- and female-biased hotspots suggest that DNA methylation in males has a dual role in driving sex biases. At PrBS with particular methylated cytosine residues, PRDM9 binding and DSB formation is favoured, whereas DNA methylation flanking the PrBS disfavours DSBs (Extended Data Fig. 9). To test this prediction, we mapped DSBs in male mice lacking functional DNMT3L, a DNA methyltransferase crucial to meiosis³². In mice with non-functional DNMT3L³³, DNA methylation is reduced by just 5–7%

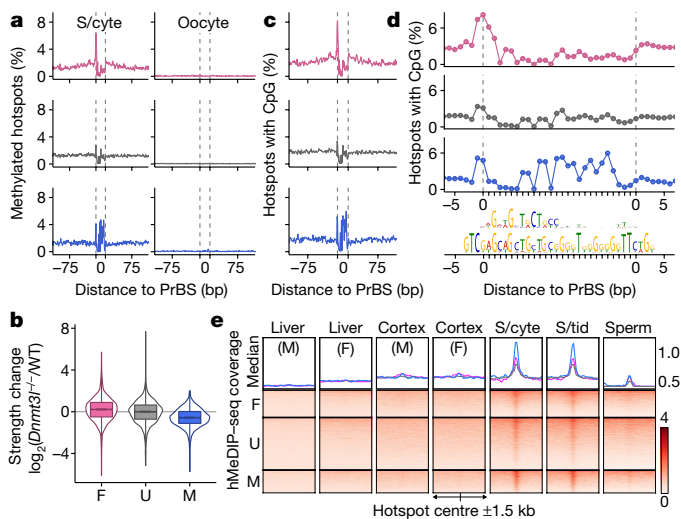


Fig. 4 | Sex-dimorphic DNA methylation at sex-biased hotspots.

Female-biased (pink), unbiased (grey) and male-biased (blue) methylated hotspots. **a**, Average methylation per base from 13 days post-partum (d.p.p.) testis⁴⁰ (spermatocytes (S/cyte); left) or 16.5 d.p.c. primordial germ cells³⁰ (oocyte; right). **b**, Sex-biased hotspots are altered in *Dnmt3l*^{-/-} males. Female-biased hotspots strengthen ($P = 10^{-123}$, Wilcoxon test) and male-biased hotspots weaken ($P = 10^{-28}$, Wilcoxon test) relative to unbiased hotspots. WT, wild type. **c**, DNA methylation mirrors CpG content. **d**, CpG density at inferred PrBS (top motif). In silico-predicted PrBS are shown below. **e**, Hydroxymethylated DNA immunoprecipitation followed by sequencing (hMeDIP-seq) at DSB hotspots; data from liver and cortex⁴¹, spermatocytes³⁸, spermatids (S/tids)³⁸ and sperm³⁸.

at PrBS (Extended Data Fig. 10a, b). Despite a reported relocation of DSBs in *Dnmt3l* mice³⁴, most SSDS-defined hotspots coincide with those in wild type ($94 \pm 2\%$; mean \pm s.d. in T1–T9). Female-biased hotspots were significantly stronger in *Dnmt3l*^{-/-} than in wild-type males, whereas male-biased hotspots were weaker (Fig. 4b, Extended Data Figs. 9, 10c, d). This strongly indicates that DNA methylation does suppress DSB formation at female-biased hotspots in males, but promotes DSB formation at male-biased hotspots. Notably, a similar pattern was seen in the *Dnmt3l*^{-/-} H3K4me3 ChIP-seq signal at hotspots (Extended Data Fig. 10e, f), suggesting that DNA methylation mediates sex differences before DSB formation.

DNA methylation occurs primarily at CpG dinucleotides³⁵, and methylation patterns at hotspots closely reflect CpG density (Fig. 4c). Thus, underlying differences in the DNA bound by PRDM9, by virtue of being frequently methylated, can result in sex biases. At female-biased hotspots, DNA methylation flanking the PrBS seems to suppress hotspot usage. Although high copy repeats are generally methylated, we find no repeat elements specifically enriched at female-biased hotspots (see Methods). Alternatively, the observed DNA methylation may favour nucleosome assembly³⁶ or exert other epigenetic effects that inhibit DSB formation at these loci in males. At male-biased hotspots, CpG density is highest at the 3' end of the PrBS (Fig. 4d; top motif). This region of the empirically determined PrBS has few apparent binding preferences, however an in silico prediction of the PrBS^{5,37} revealed a G-rich consensus at the 3' end (Fig. 4d; bottom motif), consistent with more frequent DNA methylation on the C-rich complementary strand (data not shown). This previously unappreciated complexity of PRDM9 binding seems to dictate male sex biases, and therefore, the extent of sex biases may vary among *Prdm9* alleles. The most prevalent allele of *PRDM9* in humans binds a C-rich sequence and is therefore potentially affected by DNA methylation-mediated sex biases. Male-biased recombination in humans is most prominent in distal regions¹¹, and PrBSs containing a CpG are closer than other PrBS to chromosome ends (28 ± 28 Mb and 36 ± 29 Mb, respectively; mean \pm s.d.).

Bisulfite sequencing does not distinguish between methylated (5-mC) and hydroxymethylated (5-hmC) cytosine residues. Therefore,

putatively 'methylated' nucleotides may be a compound signature of 5-mC and 5-hmC. 5-hmC is highly enriched at DSB hotspots in elutriated (primarily pachytene) spermatocytes³⁸ and in spermatids³⁸. Only residual 5-hmC at hotspots remains in sperm³⁸, and in contrast to 5-mC (Extended Data Fig. 8), the 5-hmC signal at hotspots is absent in somatic cells (Fig. 4e). Thus, 5-hmC is transiently enriched at hotspots during male meiosis. 5-hmC or DNA demethylation may therefore contribute to meiotic DSB repair, consistent with previous studies that implicated 5-HmC in the DNA damage response in mitotic cells³⁹. It remains to be seen whether the role of 5-hmC in meiotic DSB processing differs between the sexes.

Together, these data reveal extensive sex biases at the onset of meiotic recombination and several mechanisms that modulate these biases. Future examination of these mechanisms will yield further insights into how females and males differentially shape evolution of the genome.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0492-5>.

Received: 15 December 2017; Accepted: 18 July 2018;

Published online: 05 September 2018

- Lenormand, T., Engelstädter, J., Johnston, S. E., Wijnker, E. & Haag, C. R. Evolutionary mysteries in meiosis. *Phil. Trans. R. Soc. Lond. B* **371**, 20160001 (2016).
- Tortoreau, F. et al. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics* **13**, 586 (2012).
- Khil, P. P., Smagulova, F., Brick, K. M., Camerini-Otero, R. D. & Petukhova, G. V. Sensitive mapping of recombination hotspots using sequencing-based detection of ssDNA. *Genome Res.* **22**, 957–965 (2012).
- Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D. & Petukhova, G. V. Genetic recombination is directed away from functional genomic elements in mice. *Nature* **485**, 642–645 (2012).
- Smagulova, F., Brick, K., Pu, Y., Camerini-Otero, R. D. & Petukhova, G. V. The evolutionary turnover of recombination hot spots contributes to speciation in mice. *Genes Dev.* **30**, 266–280 (2016).
- Smagulova, F. et al. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* **472**, 375–378 (2011).
- Pratto, F. et al. DNA recombination. Recombination initiation maps of individual human genomes. *Science* **346**, 1256442–1256442 (2014).
- Bishop, D. K., Park, D., Xu, L. & Kleckner, N. DMC1: a meiosis-specific yeast homolog of *E. coli* recA required for recombination, synaptonemal complex formation, and cell cycle progression. *Cell* **69**, 439–456 (1992).
- Peters, H. Migration of gonocytes into the mammalian gonad and their differentiation. *Phil. Trans. R. Soc. Lond. B* **259**, 91–101 (1970).
- Kong, A. et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099–1103 (2010).
- Bhérrer, C., Campbell, C. L. & Auton, A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat. Commun.* **8**, 14994 (2017).
- Cole, F. et al. Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nat. Genet.* **46**, 1072–1080 (2014).
- Halldorsson, B. V. et al. The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.* **48**, 1377–1384 (2016).
- Lenzi, M. L. et al. Extreme heterogeneity in the molecular events leading to the establishment of chiasmata during meiosis I in human oocytes. *Am. J. Hum. Genet.* **76**, 112–127 (2005).
- Perry, J., Palmer, S., Gabriel, A. & Ashworth, A. A short pseudoautosomal region in laboratory mice. *Genome Res.* **11**, 1826–1832 (2001).
- Lange, J. et al. The landscape of mouse meiotic double-strand break formation, processing, and repair. *Cell* **167**, 695–708.e16 (2016).
- Kauppi, L. et al. Numerical constraints and feedback control of double-strand breaks in mouse meiosis. *Genes Dev.* **27**, 873–886 (2013).
- Shao, Z., Zhang, Y., Yuan, G. C., Orkin, S. H. & Waxman, D. J. MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.* **13**, R16 (2012).
- Davies, B. et al. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature* **530**, 171–176 (2016).
- Baudat, F., Imai, Y. & de Massy, B. Meiotic recombination in mammals: localization and regulation. *Nat. Rev. Genet.* **14**, 794–806 (2013).
- Narasimhan, V. M. et al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474–477 (2016).
- Walker, M. et al. Affinity-seq detects genome-wide PRDM9 binding sites and reveals the impact of prior chromatin modifications on mammalian recombination hotspot usage. *Epigenetics Chromatin* **8**, 31 (2015).

23. Tease, C. & Hultén, M. A. Inter-sex variation in synaptonemal complex lengths largely determine the different recombination rates in male and female germ cells. *Cytogenet. Genome Res.* **107**, 208–215 (2004).
24. Gruhn, J. R., Rubio, C., Broman, K. W., Hunt, P. A. & Hassold, T. Cytological studies of human meiosis: sex-specific differences in recombination originate at, or prior to, establishment of double-strand breaks. *PLoS One* **8**, e85075 (2013).
25. Liu, E. Y. et al. High-resolution sex-specific linkage maps of the mouse reveal polarized distribution of crossovers in male germline. *Genetics* **197**, 91–106 (2014).
26. Hunt, P. & Hassold, T. Female meiosis: coming unglued with age. *Curr. Biol.* **20**, R699–R702 (2010).
27. de Boer, E., Jasin, M. & Keeney, S. Local and sex-specific biases in crossover vs. noncrossover outcomes at meiotic recombination hot spots in mice. *Genes Dev.* **29**, 1721–1733 (2015).
28. de Boer, E., Stam, P., Dietrich, A. J. J., Pastink, A. & Heyting, C. Two levels of interference in mouse meiotic recombination. *Proc. Natl Acad. Sci. USA* **103**, 9607–9612 (2006).
29. Yamada, S. et al. Genomic and chromatin features shaping meiotic double-strand break formation and repair in mice. *Cell Cycle* **16**, 1870–1884 (2017).
30. Seisenberger, S. et al. The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Mol. Cell* **48**, 849–862 (2012).
31. Wang, H. et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–1688 (2012).
32. Bourc'his, D. & Bestor, T. H. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**, 96–99 (2004).
33. Vlachogiannis, G. et al. The Dnmt3L ADD domain controls cytosine methylation establishment during spermatogenesis. *Cell Rep.* **10**, 944–956 (2015).
34. Zamudio, N. et al. DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes Dev.* **29**, 1256–1270 (2015).
35. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
36. Collings, C. K., Waddell, P. J. & Anderson, J. N. Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res.* **41**, 2918–2931 (2013).
37. Persikov, A. V., Osada, R. & Singh, M. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics* **25**, 22–29 (2009).
38. Hammoud, S. S. et al. Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell Stem Cell* **15**, 239–253 (2014).
39. Kafer, G. R. et al. 5-hydroxymethylcytosine marks sites of DNA damage and promotes genome stability. *Cell Rep.* **14**, 1283–1292 (2016).
40. Jain, D. et al. *rahu* is a mutant allele of *Dnmt3c*, encoding a DNA methyltransferase homolog required for meiosis and transposon repression in the mouse male germline. *PLoS Genet.* **13**, e1006964 (2017).
41. Lin, I.-H., Chen, Y.-F. & Hsu, M.-T. Correlated 5-hydroxymethylcytosine (5hmC) and gene expression profiles underpin gene and organ-specific epigenetic regulation in adult mouse brain and liver. *PLoS One* **12**, e0170779 (2017).

Acknowledgements We thank P. Hsieh for critical feedback and the NIDDK genomics core and NHLBI flow cytometry core for assistance. This work used the computational resources of the NIH HPC Biowulf cluster (<http://hpc.nih.gov>). This research was supported by NIGMS grant R01GM084104 (G.V.P.), March of Dimes Foundation grant 1-FY13-506 (G.V.P.) and by the NIDDK Intramural Research Program (R.D.C.-O.).

Reviewer information *Nature* thanks S. Keeney, A. Pendas and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions K.B. performed in silico analyses. S.T.-S., F.S., K.-W.G.L., Y.P. and F.P. performed DMC1-SSDS experiments in male mice. F.S. performed DMC1 SSDS in females. F.P. and K.B. performed H3K4me3 ChIP-seq followed by bisulfite sequencing. K.-W.G.L., F.P. and K.B. performed sorting of ovary nuclei. G.L. performed H3K4me3 ChIP-seq in ovary. S.T.-S. performed DMC1 SSDS and H3K4me3 ChIP-seq in *Dnmt3l*^{-/-} mice. K.B. wrote the manuscript. R.D.C.-O. and G.V.P. supervised the study. All authors contributed to experimental design and critiqued the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0492-5>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0492-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.D.C.-O. or G.V.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Animal procedures. All animal procedures have been approved by the USUHS Institutional Animal Care and Use Committee or were performed according to the NIH Guide for the Care and Use of Laboratory Animals.

SSDS sample preparation and sequencing. DSBs form in the fetal ovary and the number of cells undergoing DSB repair are maximal between 15 and 16 d.p.c. (63–84% of meiotic cells are in leptotene/zygotene stages)⁹. Thus, fetal ovaries were dissected from embryos at 15.5 d.p.c. Ovaries were dissected in cold PBS and stored at -80°C until use. For SSDS, 90 or 230 ovaries were fixed in 1 ml PBS with 1% paraformaldehyde for 3 min, quenched and homogenized with a Dounce homogenizer. Cells were collected by centrifugation at 900g for 10 min using a bucket rotor. The pellet was washed in 1 ml of the following buffers: PBS, and 0.25% Triton X-100, 10 mM EDTA, 0.5 mM EGTA and 10 mM Tris, pH 8. Cells were lysed in 0.5 ml of lysis buffer (1% SDS, 10 mM EDTA, 50 mM TrisCl, pH 8 with complete protein inhibitor cocktail (Roche)) and the chromatin was sheared with Misonix sonicator with the following parameters: efficiency 1, 10 s on, 20 s off, total sonication time 4 min. Chromatin was cleared by centrifugation at 12,000g at 4°C for 10 min. The supernatant was diluted twofold by ChIP buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl, 167 mM NaCl) and dialysed against the same buffer for 5 h at 4°C .

Chromatin was incubated with 6 μg of custom-made anti-DMC1 antibody and 20 μl Dynabeads (10002D, Invitrogen) at 4°C overnight followed by washing with 500 μl of the following buffers: (1) 0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl, 150 mM NaCl; (2) 0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris-HCl pH 8, 500 mM NaCl; and (3) 0.25 M LiCl, 1% Igepal, 1 mM EDTA, 10 mM Tris-HCl, pH 8, 1% deoxycholic acid. DNA-protein complexes were eluted by two consecutive 15 min incubations at 65°C using elution buffer (0.1 M NaHCO_3 , 1% SDS, 5 mM dithiothreitol (DTT)). The eluates were combined and crosslinking was reversed at 65°C for 5 h. The samples were deproteinized and cleaned up with MinElute PCR purification kit (QIAGEN).

The sequencing library was prepared as previously described⁴ with minor modifications. In brief, the end repair step was done in $1 \times$ T4 DNA ligase buffer with 10 mM ATP in the presence of 0.25 mM dNTPs, 0.6 U T4 DNA polymerase, 0.5 U Klenow, 2 U T4 polynucleotide kinase for 30 min at 20°C . DNA was purified with MinElute kit. The second step was done in the presence of 1 mM dATP and 1 U Klenow Exo-. Reaction was incubated at 37°C for 30 min and DNA was purified with MinElute kit. To enrich for ssDNA the sample was denatured for 2 min at 95°C , then cooled to room temperature. The sequencing adaptor mix (Illumina) was diluted 1:200 and added for the ligation step. DNA was purified by MinElute kit and amplified using Phusion Polymerase (0.5 μl per reaction), in the presence of 1 μl of each Illumina PE primers. The following parameters were used: initial denaturation at 98°C for 30 s; 21 cycles: 98°C 10 s, 65°C 30 s, 72°C 30 s; final extension for 5 min at 72°C . Size selection was done in 2% agarose gel, 180–250 bp slice was excised and purified using MinElute Gel Extraction kit.

SSDS was performed from adult whole testis using a protocol described previously⁴².

H3K4me3 ChIP-seq sample preparation and sequencing. For H3K4me3 ChIP-seq in oocytes, we isolated SCP3-positive meiotic oocytes from 14 15.5-d.p.c. females using batch isolation of tissue-specific chromatin for immunoprecipitation (BiTS-ChIP). In brief, this method uses FACS to isolate nuclei on the basis of the presence of an intra-nuclear marker (in this case, anti-SCP3; Santa Cruz sc-74569). Oocytes were isolated by gating for 4C nuclei with SCP3 signal above that from secondary antibodies alone. The gating strategy is outlined in Supplementary Fig. 1. The KAPA Hyper Prep kit (KR0961) was used to prepare the sequencing library due to the limited starting material relative to experiments in whole testis.

Testis sample preparation was performed as described previously⁴. The following antibodies were used: anti-DMC1 (Santa Cruz C-20, sc-8973), anti-DMC1 (custom-made), and anti-H3K4me3 (Millipore 07-473 or Abcam 8580). All sequencing was performed on an Illumina HiSeq 2500 at the NIDDK genomics core.

Targeted bisulfite sequencing. H3K4me3 ChIP-seq was performed from whole testis as described above; however, the library preparation protocol was modified to perform bisulfite sequencing. Immediately after sequencing adaptor ligation, we performed bisulfite conversion using the Qiagen EpiTect Bisulphite Kit (59104). Subsequent library amplification was performed using KAPA HiFi Uracil+ Kit, which is designed to tolerate uracil residues in bisulfite-treated DNA. Sequencing was performed on an Illumina HiSeq 2500 at the NIDDK genomics core and on an Illumina HiSeq X Ten by Admera Health.

Alignment of sequencing reads. For SSDS, reads were aligned to the genome and ssDNA-derived reads were identified using the SSDS processing pipeline³. In brief, the first read of each mate pair is mapped to the genome with bwa (v.0.7.12)⁴³. The

second read is then mapped to the genome using a modified bwa algorithm that finds the longest mapping suffix for each read. ssDNA is determined from the structure of inverted terminal repeats on the first and second end reads. The SSDS alignment pipeline is available at <https://github.com/kevbrick/callHotspotsSSDS>.

All other sequencing data were aligned to the reference genome using bwa aln (0.7.12)⁴³.

Evaluation of SSDS sensitivity for DSB detection in ovary. To estimate the lower hotspot detection limit using SSDS, we generated a DSB map using 2×10^5 testis cells from *Hop2*^{-/-} (also known as *Psmc3ip*^{-/-}) mice⁴⁴. HOP2 is required for DSB repair, and approximately 37% of cells in testis of *Hop2*^{-/-} mice harbour unrepaired DSBs (data not shown). The signal percentage of tags (SPoT) for this sample was 21% (sample N1; Extended Data Fig. 1a); this is slightly lower than the SPoT for DSB maps from whole testis (23–46%; samples T1–T9; Extended Data Fig. 1a), but far above the 2% SPoT expected by chance. The estimated library size of N1 was about $10 \times$ smaller than for the smallest whole testis sample (Extended Data Fig. 1d), probably because of the limited starting amount of DNA. Because small library size complicates hotspot detection, we first attempted to map DSBs in females using approximately $10 \times$ more target cells (a pool of 90 ovaries; approximately 9×10^5 target cells). This sample (O2) had a library size close to that of wild-type whole testis samples (Extended Data Fig. 1d) but a lower SPoT (7%; Extended Data Fig. 1a), suggesting that hotspot DNA recovery from oocytes was less efficient than from testis. Subsequently, we pooled 230 fetal ovaries to generate a second ovary-derived DSB map (O1). This map was of similar quality to DSB maps derived from testis (SPoT = 33%; library size = 3.8×10^7 fragments; sample O1; Fig. 1, Extended Data Fig. 1a, d).

DSB hotspot identification. Uniquely mapping fragments unambiguously derived from ssDNA (ssDNA type 1) and having both reads with a mapping quality score ≥ 30 were used for identifying hotspot locations (peak calling). NCIS⁴⁵ was used to estimate the background fraction for each library. Peak calling was performed using MACS (v.2.1.0.20150420)⁴⁶ with the following parameters: -ratio [output from NCIS] -g mm -bw 1000 -keep-dup all -slocal 5000. We use a mixture-model-based approach that accounts for GC-biases to calculate a corrected *P* value for each hotspot (model = negative binomial; number of iterations for refinement = 100)⁴⁷. *P* values were adjusted for multiple testing using the Benjamini–Hochberg method. Hotspots with a GC-corrected *P* > 0.05 and DSB hotspots within regions previously blacklisted⁵ were discarded.

H3K4me3 peak calling. Uniquely mapping reads with a mapping quality score ≥ 30 were used for peak calling. NCIS was used to estimate the background fraction relative to an input DNA library. Peak calling was performed using MACS (v.2.1.0.20150420)⁴⁶ with the following parameters: -ratio [output from NCIS] -g mm -bw 1000 -keep-dup all -slocal 5000. Peak strength was subsequently calculated by subtracting the NCIS⁴⁵ normalized input read count from the ChIP-seq read count.

H3K4me3 peaks that result from PRDM9 activity were inferred from the overlap with DSB hotspot locations. We also exclude any such sites that coincide with a DSB hotspot in *Prdm9*^{-/-} mice⁴, as these are sites of PRDM9-independent H3K4me3.

Hotspot overlaps and merging hotspot sets. Unless otherwise stated, when assessing whether hotspots occur at the same location, we restrict the overlap to the ± 200 bp region of DSB hotspots. Previously, we have shown that using a ± 200 bp regions is sufficient for detecting true overlaps and limits the number of spurious overlaps⁴. We merged DSB hotspots from the best testis and ovary samples (T1 and O1, respectively). The centre of overlapping hotspots was defined as the mean centre point of the T1 and O1 hotspot and the flanks were defined as the maximum distance from this centre to the original T1 and O1 hotspot edges. Non-overlapping hotspots from each sample were retained as originally defined.

Strength metrics at hotspots. Hotspot strength measured by SSDS was calculated as described previously⁵. In brief, the centre-point of the Watson- and Crick-strand ssDNA fragment distributions was used to define the hotspot centre. ssDNA fragments on the Watson (top) strand to the left and Crick (bottom) strand to the right of this centre were considered signal. Background was extrapolated from the count of ssDNA fragments of opposite polarity around the centre (excluding the very centre of the hotspot). Hotspot strength is then calculated as the signal minus the background fragment count. This strength is used as a proxy for DSB frequency^{4,5}. Scripts for peak calling and strength quantitation are available at <https://github.com/kevbrick/SSDSpipeline>. Hotspot strength measured from SPO11-oligo mapping was calculated as the sum of the strengths of SPO11 oligo-nucleotide peaks overlapping each hotspot. SPO11-oligo peaks were downloaded from the processed data associated with the Gene Expression Omnibus (GEO) record (GSM2247727)¹⁶. H3K4me3 strength at hotspots was calculated as the sum of the strength of overlapping H3K4me3 peaks.

Sample SPoT reduction. SPoT is calculated as the number of in-hotspots ssDNA fragments divided by the total number of ssDNA fragments. To reduce SPoT, the required number of randomly selected in-hotspot fragments is discarded. If necessary, background fragments are added from an input DNA library.

Default hotspots. Hotspots found in *Prdm9*^{-/-} DSB maps and lacking a putative PrBS were designated as 'default' hotspots. Default hotspots constitute 8.5% (427 out of 5,021) of female-biased hotspots but only 1.3% (58 out of 4,169) of male-biased hotspots, therefore we distinguish PRDM9-dependent and default hotspots in subsequent analyses.

Sex bias determination. We used MANorm¹⁸ to infer differential usage of DSB hotspots between the T1 and O1 samples. All hotspots from the T1, O1 merge were used as 'common' peaks. Hotspots on chromosomes X, Y and M were excluded from this analysis. MANorm *P* values were adjusted for multiple testing using the Benjamini–Hochberg method and hotspots with a corrected *P* < 0.01 were considered differential.

Cluster analysis. We identified groups of adjacent hotspots that shared the same sex-bias (female-biased, male-biased or unbiased). Only uninterrupted runs of hotspots with the same bias were considered. A cluster is >1 consecutive hotspots with the same sex bias. To estimate the expected numbers of hotspots in clusters, hotspot bias designations were shuffled and the aforementioned process was repeated. Approximately 10,000 iterations of this randomization process were performed. *P* values are calculated from the empirical distribution of expected values for clusters of each size.

Generation of randomized maps of DSB hotspots. DSB hotspot locations were randomized as described in⁴. In brief, hotspots were uniformly distributed per chromosome, but prohibited from being placed at unmappable regions. Specifically, a mappability score for 40-bp sequencing reads at each mm10 base was calculated using the GEM library (20100419-003425)⁴⁸. Hotspot excluded regions were defined as annotated assembly gaps from UCSC, in addition to 1 kb genomic intervals with <50% uniquely mappable bases plus 1 kb either side. Hotspot width and strength were preserved at the randomized location for each hotspot.

Genetic crossovers. The locations of mouse crossovers were obtained from the Collaborative Cross²⁵. To assess whether the increased male crossover rate in the q-arm subtelomeric region is *Prdm9* dependent, we inferred genetic crossovers that were likely to have been formed by the B6, CAST or PWD alleles of *Prdm9*. Crossovers likely defined by the B6, CAST and PWD/PWK alleles of *Prdm9* were determined by first identifying crossovers that occurred in hybrids involving any of these pure strains (MGP, MGM, PGP, PGM). Crossovers that coincided with a single DSB hotspot from only one of the parental *Prdm9* alleles were then designated as having originated from that allele. All hotspots between B6, 129 and WSB mice that coincided with a B6-, C3H- or 13R-defined DSB hotspot were designated as having originated from *Mus musculus domesticus*. Crossovers that did not overlap any DSB hotspot from either parental genotype and those from crosses for which DSB hotspot maps were not available were designated as 'ambiguous'. Crossovers derived from all *Prdm9* alleles showed similar sub-telomeric enrichment for males relative to females (not shown), consistent with previous reports²⁵. Thus, this enrichment appears a *Prdm9*-independent phenomenon.

DNA methylation. For bisulfite sequencing (BS) samples BS_{SC}, BS_{H3K4me3}, BS_{SG}, BS_{DAA}, BS_{DWT} (see Extended Data Fig. 1h, i), bismark⁴⁹ was used to align whole-genome bisulfite sequencing data to the reference mm10 genome. For BS_{rahu}, BS_{PGC}, BS_{I3.5}, BS_{LIV}, BS_{MG} (see Extended Data Fig. 1i), pre-processed nucleotide-resolution methylation data were available, therefore UCSC liftover⁵⁰ was used to convert these data from mouse mm9 to mouse mm10 genomic coordinates where necessary. hMeDIP-seq reads³⁸ were mapped to the genome using bwa mem (0.7.12)⁴³ and default parameters.

To examine methylation patterns, we first inferred a high confidence set of PrBSs at DSB hotspots. We used FIMO⁵¹ to identify matches to the B6 PrBS⁴ position weight matrix (PWM) within hotspots. The number of matches to the PWM of the PrBS depends on the alignment score threshold used, therefore we identified the PWM *P* value alignment threshold that yielded the maximal number of DSB

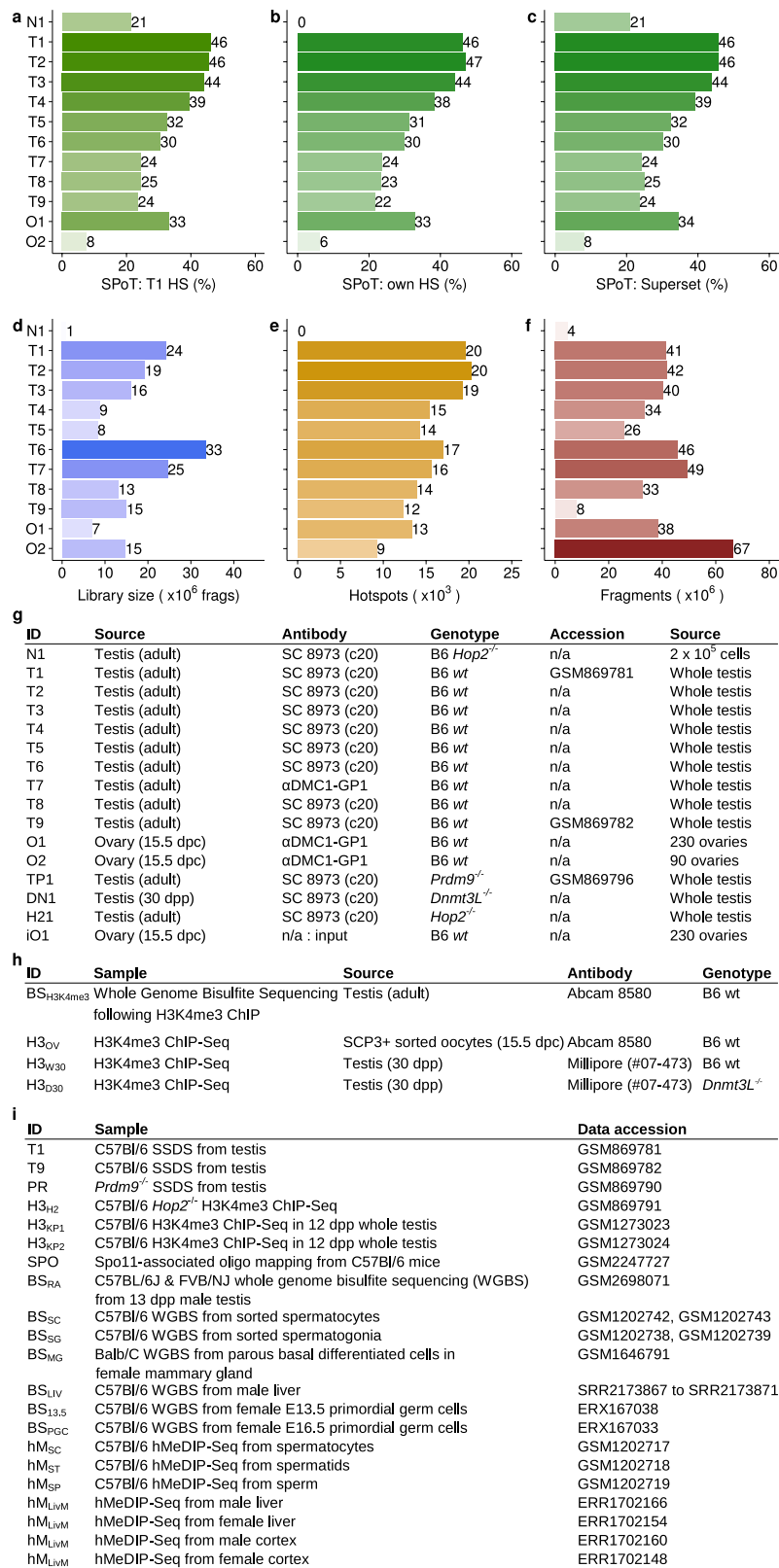
hotspots with a single match to the PRDM9 PWM within the central ± 250 bp. We tested the following threshold values: $P \leq 5 \times 10^{-3}$, 1×10^{-4} , 2×10^{-4} , 4×10^{-4} , 6×10^{-4} , 8×10^{-4} , 1×10^{-5} , 2×10^{-5} , 4×10^{-5} , 6×10^{-5} , 8×10^{-5} , 1×10^{-6} , 1×10^{-7} , 1×10^{-8} . 12,097 / 19,053 hotspots (63%) contained a single PrBS at the optimal threshold ($P \leq 4 \times 10^{-4}$).

High copy repeats at DSB hotspots. High copy repeats determined by RepeatMasker (<http://www.repeatmasker.org/>, accessed 11 December 2017) for the mouse mm10 genome were split into 67 groups by family. Repeat families that overlapped <0.2% of any hotspot set were excluded. To assess whether hotspots biased to each sex associated with different DNA high copy repeat families, we counted the frequency of repeats overlapping the central ± 200 bp of female-biased, unbiased and male-biased hotspots. To assess differences, we performed a two-sided binomial test for all pairwise comparisons (male/female, male/unbiased, female/unbiased). *P* values were Bonferroni corrected to account for multiple testing and a corrected *P* < 0.01 was used to assess differences. Repeats that showed significantly different enrichment in one set of hotspots compared to both others were investigated. From this analysis, two families of long terminal repeat (LTR) retrotransposons (LTR and LTR/ERV) are depleted at male-biased hotspots, whereas LINE/L1, SINE/Alu and SINE/1D elements are depleted at female-biased hotspots relative to the other two sets. Notably, however, no repeat families are increased exclusively at either set of sex-biased hotspots.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper

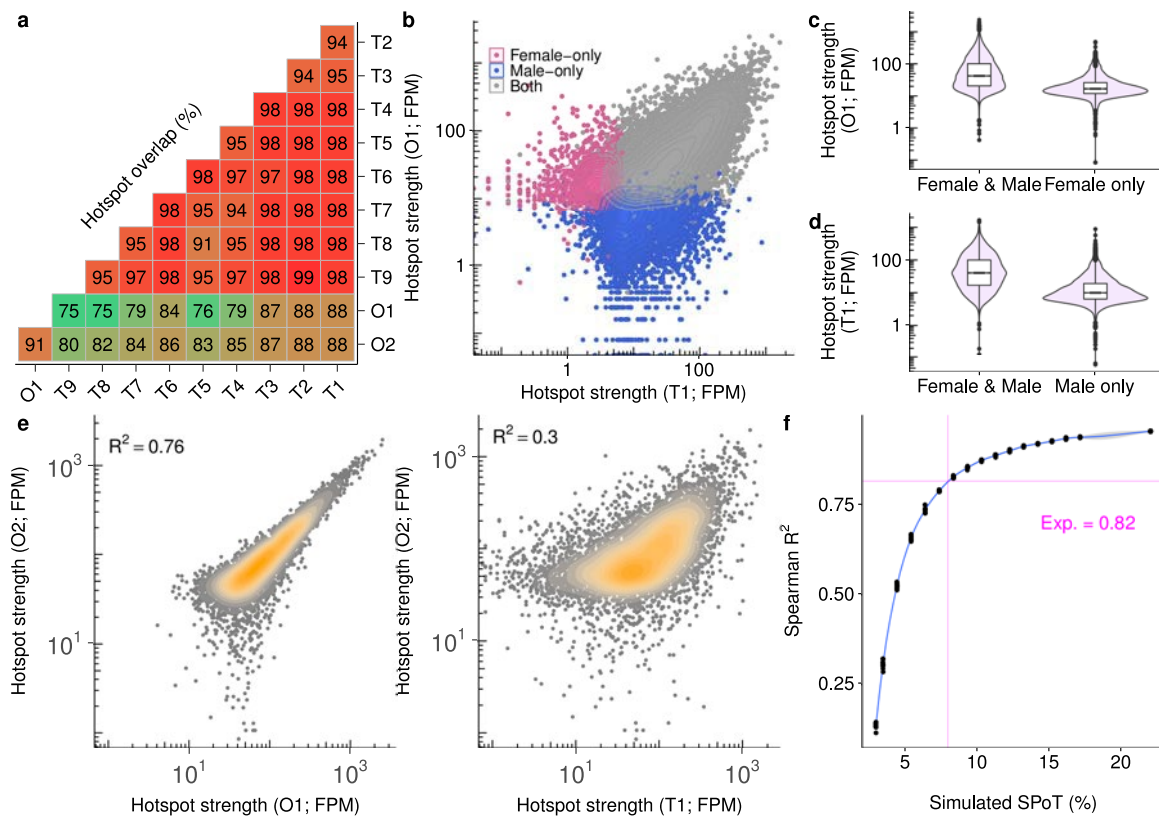
Data availability. Sequencing data are archived at the Gene Expression Omnibus (GEO) under accession GSE99921.

42. Brick, K., Pratto, F., Sun, C.-Y., Camerini-Otero, R. D. & Petukhova, G. Analysis of meiotic double-strand break initiation in mammals. *Methods Enzymol.* **601**, 391–318 (2018).
43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
44. Petukhova, G. V., Romanienko, P. J. & Camerini-Otero, R. D. The Hop2 protein has a direct role in promoting interhomolog interactions during mouse meiosis. *Dev. Cell* **5**, 927–936 (2003).
45. Liang, K. & Keleş, S. Normalization of ChIP-seq data with control. *BMC Bioinformatics* **13**, 199 (2012).
46. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
47. Teng, M. & Irizarry, R. A. Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data. *Genome Res.* **27**, 1930–1938 (2017).
48. Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS One* **7**, e30377 (2012).
49. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
50. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
51. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
52. Baker, C. L., Walker, M., Kajita, S., Petkov, P. M. & Paigen, K. PRDM9 binding organizes hotspot nucleosomes and limits Holliday junction migration. *Genome Res.* **24**, 724–732 (2014).
53. Oey, H., Isbel, L., Hickey, P., Ebaid, B. & Whitelaw, E. Genetic and epigenetic variation among inbred mouse littermates: identification of inter-individual differentially methylated regions. *Epigenetics Chromatin* **8**, 54 (2015).
54. dos Santos, C. O., Dolzhenko, E., Hodges, E., Smith, A. D. & Hannon, G. J. An epigenetic memory of pregnancy in the mouse mammary gland. *Cell Rep.* **11**, 1102–1109 (2015).
55. Bonn, S. et al. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44**, 148–156 (2012).



Extended Data Fig. 1 | Sample details and quality metrics for DSB maps. **a**, The signal portion of tags is calculated for all samples at hotspots (HS) identified in the T1 sample. Sample identifiers are in panel **g**. **b**, **c**, Hotspots identified in each respective sample (**b**) or hotspots in the combined T1/O1 superset (**c**). Peak calling was not performed for N1 (see Methods). **d**, The estimated library size (x) was inferred using bisection root finding for $f(x) = (1 - N_{NR}/x) - \exp(N_{tot}/x)$, $10^4 \leq x \leq 10^{12}$. N_{NR} , number of unique fragments; N_{tot} = total number of fragments. **e**, The number of hotspots identified in each sample. **f**, The number of

ssDNA fragments sequenced for each sample. **g**, Details of SSDS samples. Sample N1 was generated from *Hop2*^{-/-} mice using 2×10^5 cells. This sample was run in single-end mode, and not processed through the ssDNA pipeline. Previously published samples are referenced by the GEO accession number. Note that the use of the SC 8973 (c20) and anti-DMC1-GP1 antibodies gave indistinguishable SSDS results in males (Fig. 2c, Extended Data Fig. 2a). **h**, Details of samples from H3K4me3 ChIP-seq. **i**, Details of publicly available datasets used.

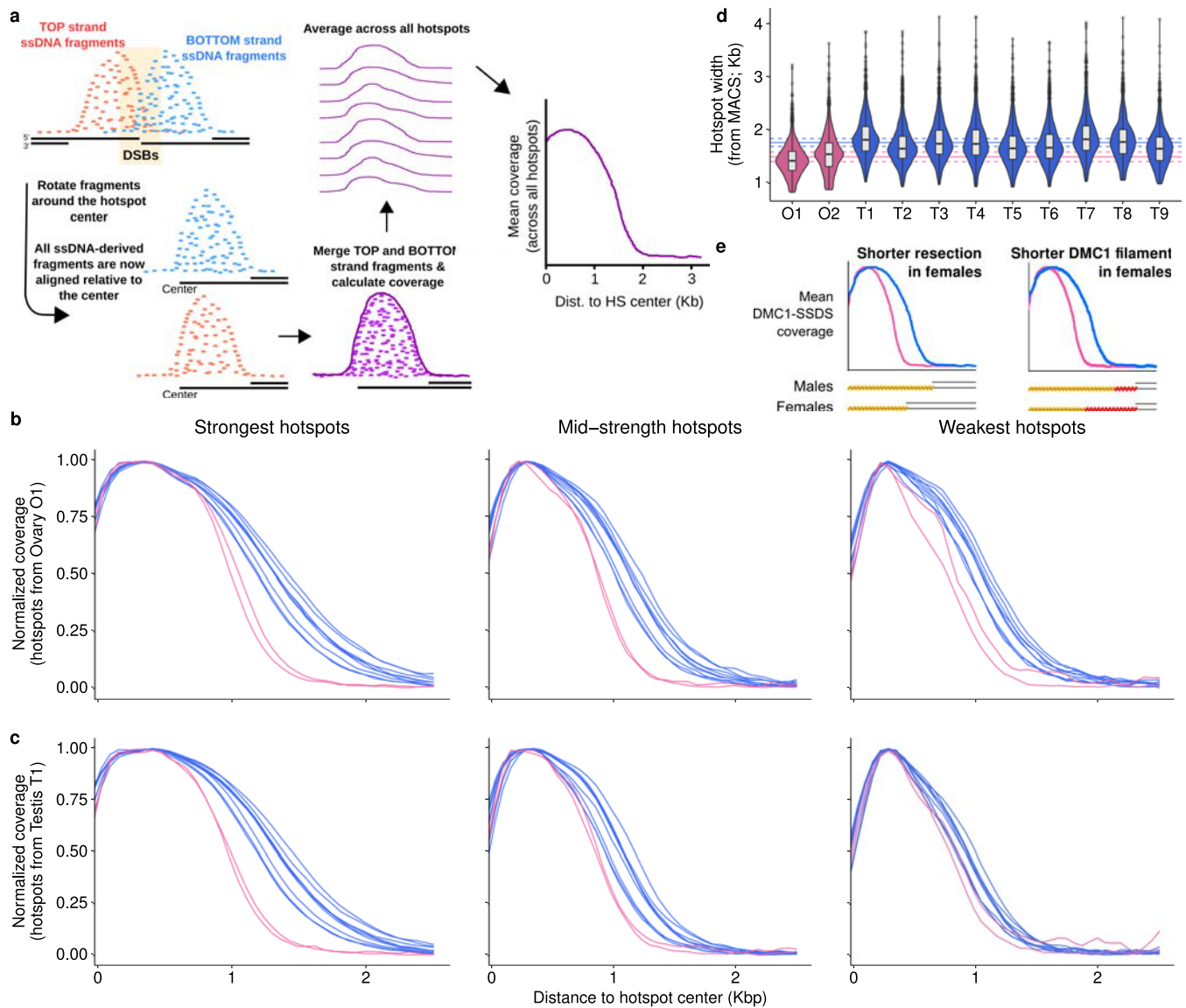


Extended Data Fig. 2 | Most DSB hotspots are used in both male and female meiosis.

a, The maximum reciprocal overlap between hotspots in each sample was calculated using the central ± 200 bp of hotspots.

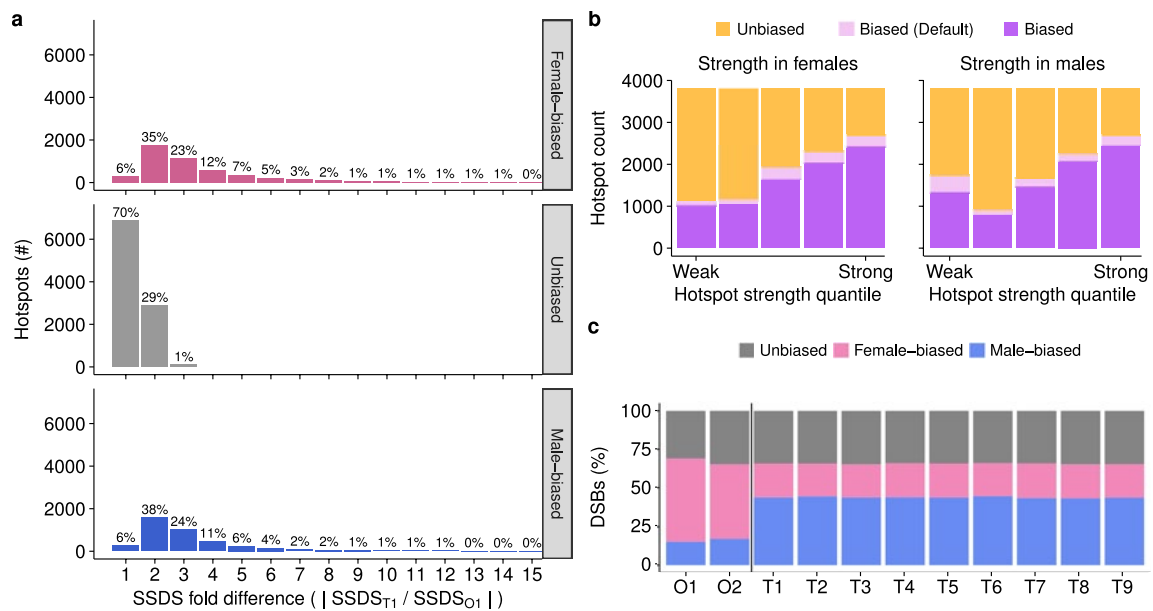
b, Hotspots exclusively found in either sex are weak. Hotspots were split into those found in both O1 and T1 (both; grey), O1 but not T1 (female-only; pink) and T1 but not O1 (male-only; blue). **c**, Female-only hotspots are weak in females, relative to shared hotspots. **d**, Male-only hotspots are weak in males, relative to shared hotspots. **e**, The O2 SSDS correlates better with hotspot strength in ovary (O1) than in testis (T1). Only hotspots that are detected in both samples are shown for each comparison. Note that the correlation between hotspot strength in ovary samples (Spearman's $R^2_{O2O1} = 0.76$) is not as high as that between replicates of SSDS in males (minimum Spearman's $R^2 \geq 0.9$; Fig. 2). **f**, Noise in SSDS estimates can fully explain this diminished correlation between ovary-derived SSDS maps. We generated a series of downsampled O1 SSDS datasets to test

whether reducing the SPoT value would reduce the maximum possible R^2 value. For each simulated dataset, signal reads were randomly chosen from the uniquely mapping in-hotspot ssDNA fragments of the O1 DMC1 SSDS sample. Background fragments were randomly chosen from all uniquely mapping ssDNA fragments of the O1 input DNA SSDS sample. Samples at different SPoTs were then generated by varying the number of signal and background-derived reads ($SPoT = \text{signal}/(\text{signal} + \text{background})$). The number of fragments was matched to the number of uniquely mapping fragments in O2. Ten replicate samples were generated for each SPoT, and the correlation coefficient (Spearman's R^2) with the original O1 SSDS sample was calculated. The magenta lines indicate the expected maximum R^2 for a sample with a SPoT matching that of O2. The expected maximum R^2 is very close to the observed R^2 . Thus, noise in SSDS estimates can reduce the R^2 to within the observed range for a sample of this quality.



Extended Data Fig. 3 | SSDS signal at hotspots is narrower in ovaries than in spermatocytes. **a**, SSDS coverage is a measure of DMC1-bound ssDNA either side of each meiotic DSB. In a population of meocytes, DSBs will occur in a several hundred nucleotide window around the hotspot centre (orange rectangle). To assess coverage, we first convert the position of each SSDS fragment into the distance along the ssDNA from the hotspot centre. Merging the top and bottom strand fragments in this way increases coverage twofold and minimizes the influence of asymmetric gaps and fluctuations in coverage. Coverage at each hotspot was normalized by the maximum value at the hotspot to prevent strong hotspots from dominating the average profile. The average normalized coverage across all hotspots was then calculated. **b**, **c**, DSB hotspots identified in females (ovary sample O1) (**b**) and males (testis sample T1) (**c**) were each split into three bins by strength. Coverage was calculated for all nine male and two female samples for each set. The SSDS signal is narrower for all female samples compared to male samples. The difference is particularly pronounced at stronger hotspots, in which coverage estimates are most accurate. At the widest point, the mean male and female profiles diverge by approximately 0.4 kb. **d**, We also examined

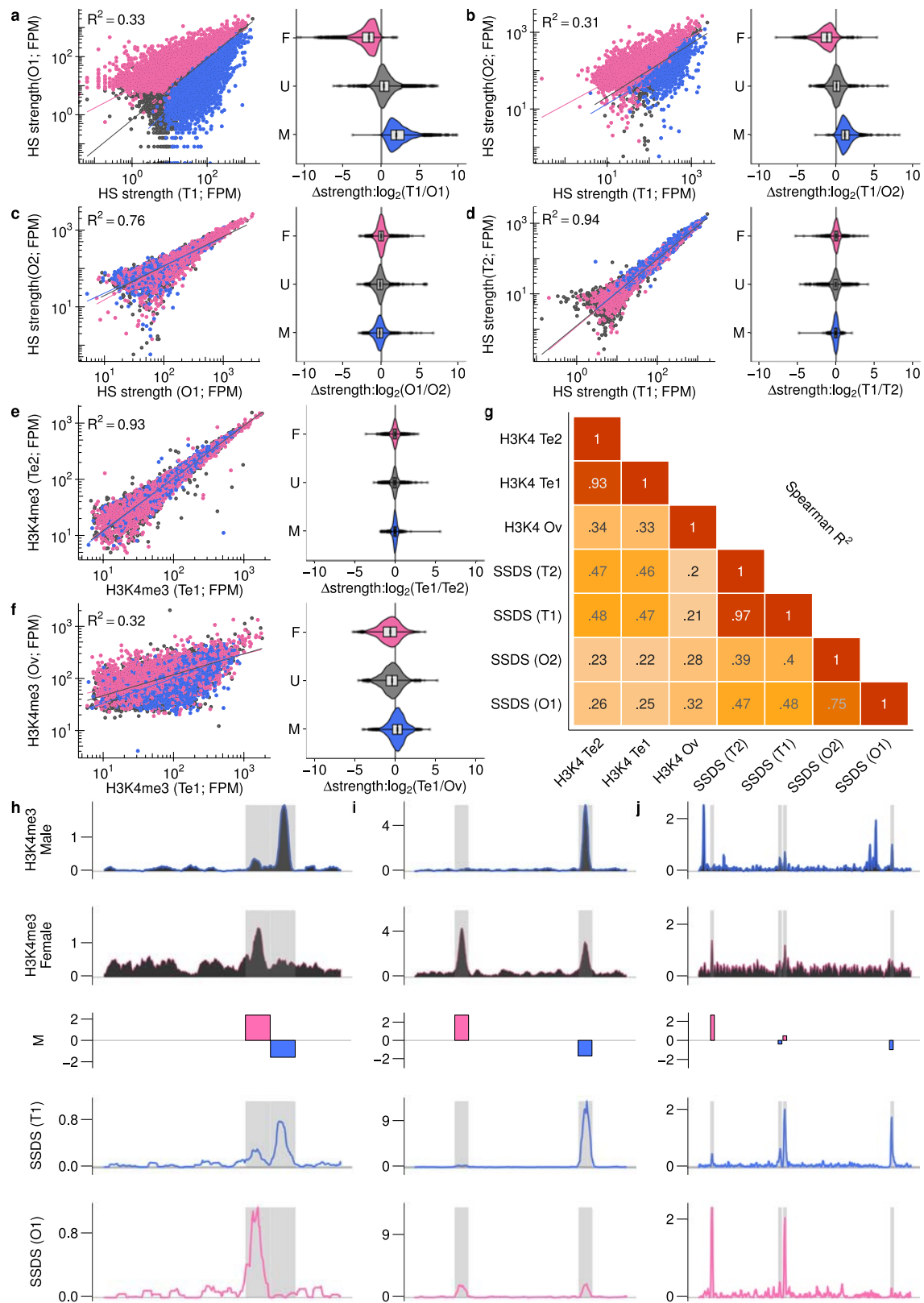
the model-based analysis of ChIP-seq (MACS)-determined hotspot boundaries to further negate the possibility that the average profiles in **b** and **c** are not a reflection of the population. By this metric, the mean hotspot width estimated from male samples ($1,759 \pm 73$ bp; mean (solid blue line) \pm s.e.m. (dashed blue lines); $n = 9$) is significantly wider than the mean width of hotspots in female samples ($1,490 \pm 89$ bp; mean (solid pink line) \pm s.e.m. (dashed pink lines); $n = 2$) ($P = 0.0007$; t -test). Because sequencing quality and sample SPoT can affect width estimates, we processed each sample as follows: we reduced the SPoT of each sample to that of the lowest quality sample (O2; see Methods), considering only uniquely mapping and high quality ($Q > 30$) ssDNA type 1 fragments. We then reduced all samples to have the same number of fragments as the smallest. On these datasets, we performed peak calling and retained only DSB hotspots that were called in all samples ($n = 1,975$). **e**, Potential mechanistic explanations for the difference in SSDS signal between males and females. These differences may manifest in all meocytes or in sub-populations. Notably, we see no evidence of shape differences at hotspots in sub-populations of spermatocytes (data not shown).



Extended Data Fig. 4 | Most meiotic DSBs occur at sex-biased hotspots.

a, Quantification of SSDS fold change at sex biased and unbiased hotspots. The percentages show the percentage of hotspots in each category with a given absolute fold change. **b**, The hotspots in the testis/ovary superset were split into quintiles by strength in either females (left) or males (right). In both sexes, over 60% of the strongest hotspot subset exhibit sex-biased DSB formation. This is a proxy for the true amount of sex-biased DSB formation. In progressively weaker hotspot sets, fewer biased hotspots are detected. One outlier is the set of weak male hotspots. This set contains

many female-biased default hotspots that form independently of PRDM9. **c**, We quantified the total in-hotspot SSDS signal at female-biased, unbiased and male-biased hotspots in the two ovary-derived samples and in the nine testis samples. In all cases, over half of the in-hotspot sequencing tags (referred to as total DSBs) occur at sex-biased hotspots. Hotspots biased towards usage in females are enriched in ovary samples, while those biased towards male usage are enriched in testis-derived samples.

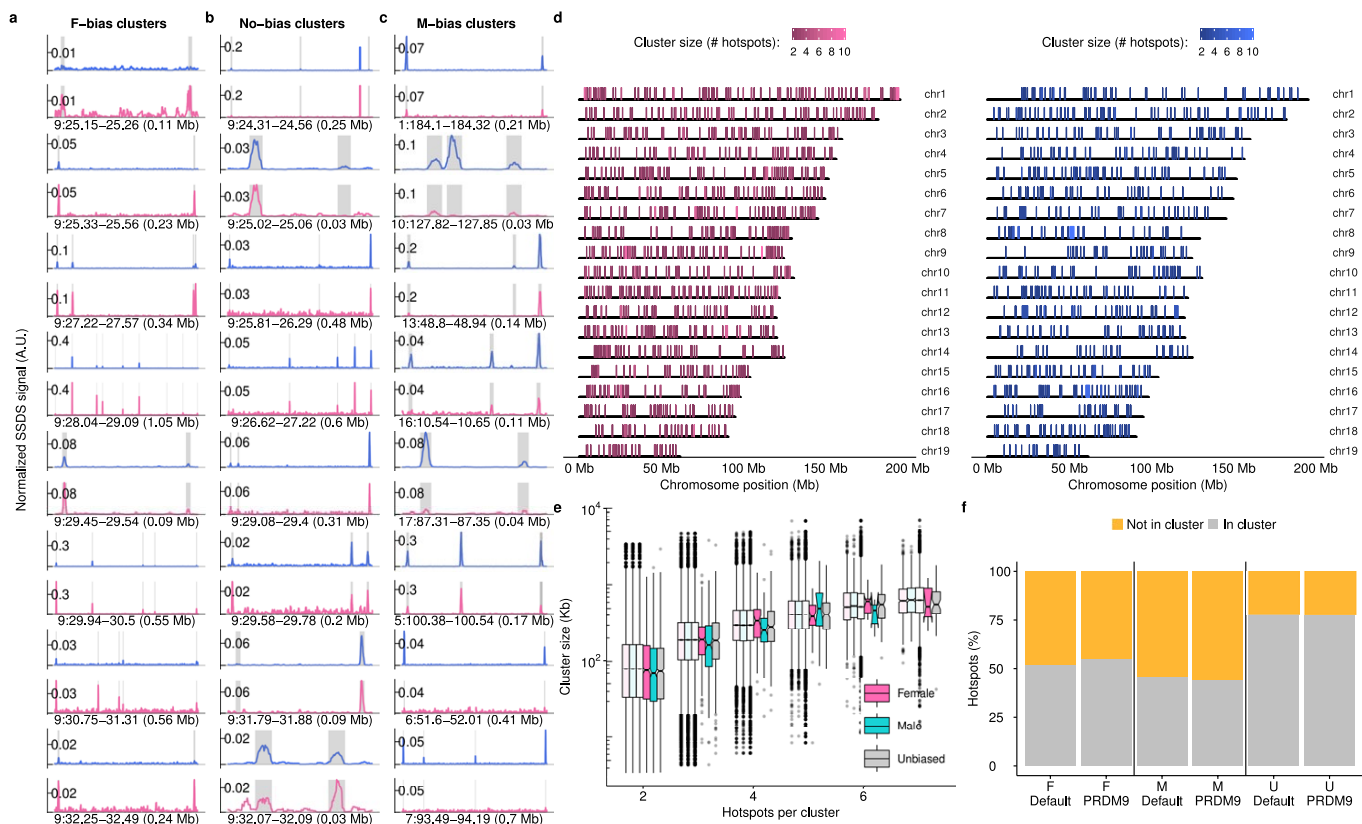


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Sex biased hotspots are consistent across replicates and are defined before DSB formation. Hotspot strength was calculated at all autosomal hotspots from the merged O1/T1 DSB maps. The strength of hotspots was re-calculated in two testis (T1 and T2) and two ovary (O1 and O2) maps. Female-biased (pink), unbiased (grey) and male-biased (blue) hotspots were determined by comparing the T1 and O1 maps. These hotspots are coloured the same in all panels. **a**, Sex-biased hotspots are distributed as expected when comparing the O1 and T1 DSB maps. These data are also plotted in Fig. 2d, e. **b**, Sex-biased hotspots exhibit the same sex-biases in the O2 sample. **c**, **d**, Sex-biased hotspots exhibit no biased usage between samples derived from mice of the same sex. **e–g**, Sex biases that precede DSB formation were studied by performing H3K4me3 ChIP-seq in FACS-purified fetal oocytes at 15.5 d.p.c. (see Methods). The H3K4me3 signal at hotspots was quantified and compared to existing maps of H3K4me3 in juvenile mouse testis⁵². **e**, The H3K4me3 signal at hotspots is tightly correlated in replicate samples from

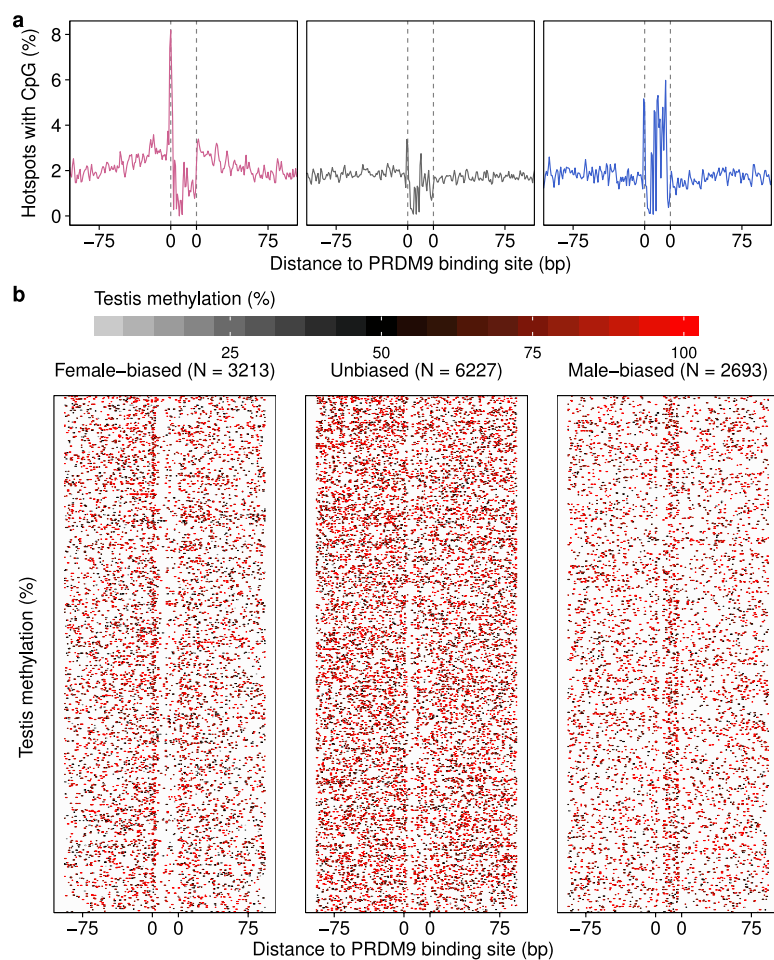
mouse testis ($Te1 = H3_{KP1}$, $Te2 = H3_{KP2}$; Extended Data Fig. 1i). **f**, Similar to what we observe when examining the SSDS signal at DSB hotspots, there is extensive variation in the H3K4me3 signal at hotspots between male and female meiosis. This indicates that sex biases are established before DSB formation. Sex biases determined using SSDS remain broadly conserved when we compare H3K4me3 in females to males. **g**, H3K4me3 at hotspots is better correlated with SSDS from the respective sex.

h–j, Sex biases in H3K4me3 ChIP-seq parallel the differences in the SSDS signal. H3K4me3 ChIP-seq coverage is shown in the top panels; testis ($H3_{KP1}$; blue) and ovary ($H3_{OV}$; pink). The middle panel shows the \log_2 fold difference (M) between the SSDS signal in testis (T1) and ovary (O1). SSDS coverage for these samples is shown in the bottom panels. Grey boxes represent DSB hotspot positions. To allow for quantitative cross comparison, the coverage in each sample is normalized by the median signal strength at DSB hotspots in that sample. The genomic coordinates of each window are given underneath.



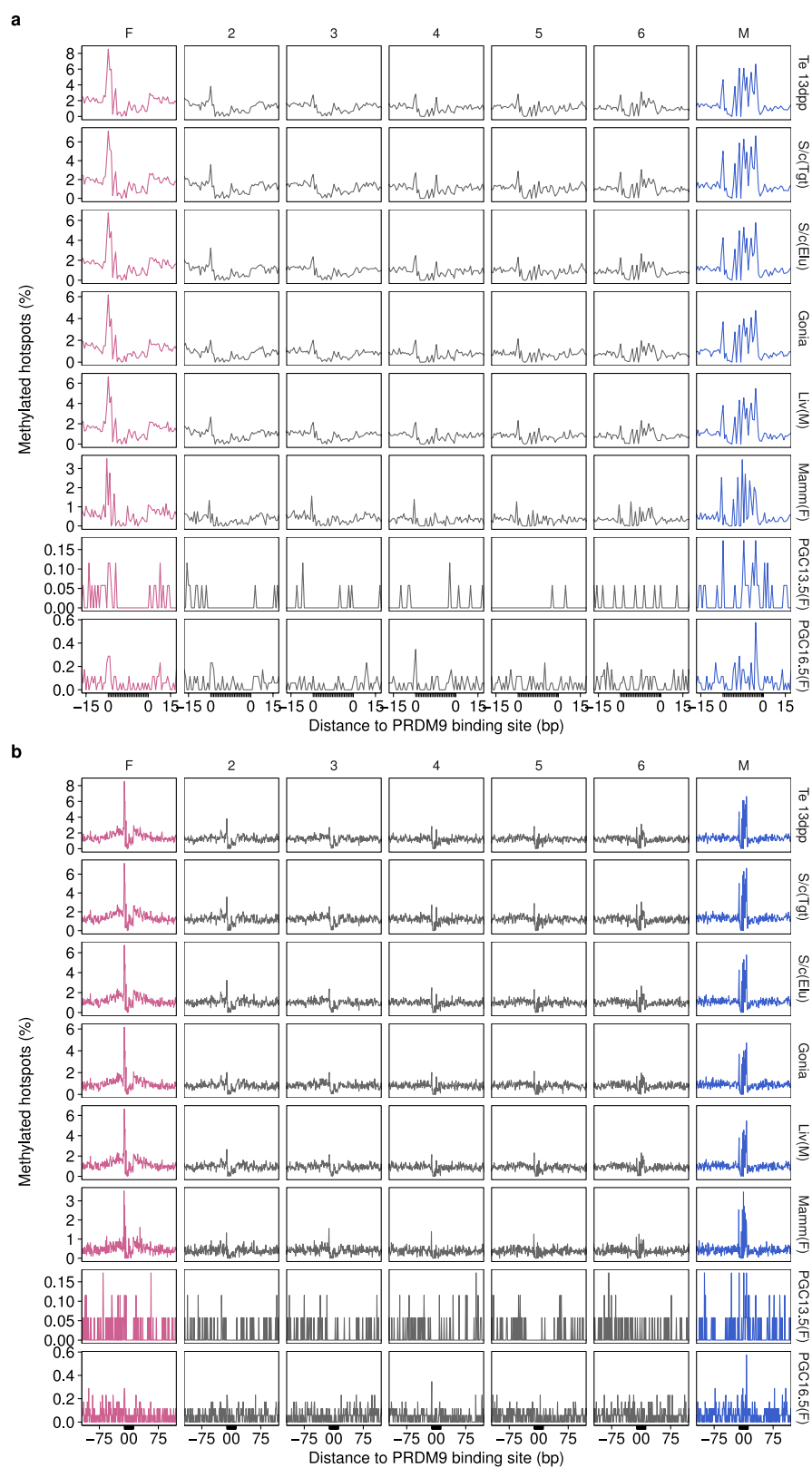
Extended Data Fig. 6 | Clustering of sex-biased hotspots. a–c, SSDS coverage at a subset of biased hotspot clusters. The female-biased (**a**) and unbiased (**b**) clusters are those shown in Fig. 3a. **c**, Because no male-biased clusters are depicted in Fig. 3a, eight clusters were randomly chosen. SSDS coverage for testis (T1; blue) and ovary (O1; pink) are shown for each cluster. To allow for quantitative cross comparison, coverage in each sample is normalized by the median hotspot strength. Grey boxes represent DSB hotspot positions. **d**, Genomic patterning of sex-biased DSB hotspots. Female-biased (left; pink) and male-biased (right; blue) hotspot clusters on all autosomes. Biased hotspots do not exhibit particular spatial patterning, aside from a slight enrichment of female-biased hotspots at

the q-arm telomere. **e**, The physical size of hotspot clusters scales with the number of hotspots per cluster. It therefore seems unlikely that clustering results from a physical size constraint imposed by sex-specific chromatin structure. Notably, however, the presence of such a size constraint may be masked by the presence of a large number of clusters that occur by chance. Semi-transparent box plots show the expected size distribution for randomly distributed clusters ($n = 1,000$ bootstraps). Clusters of three male-biased hotspots are marginally smaller than expected. There are no significant differences for clusters of other sizes. **f**, Similar proportions of PRDM9-defined and default hotspots occur in clusters. Hotspots in clusters of ≥ 2 consecutive hotspots of the same type were counted.



Extended Data Fig. 7 | Differing patterns of DNA methylation at sex-biased hotspots. **a**, Mean DNA methylation⁴⁰ at the putative PrBS (grey bar) of female-biased (pink), unbiased (grey) and male-biased (blue) hotspots. Note that this panel is also shown in Fig. 4a. **b**, Heat map

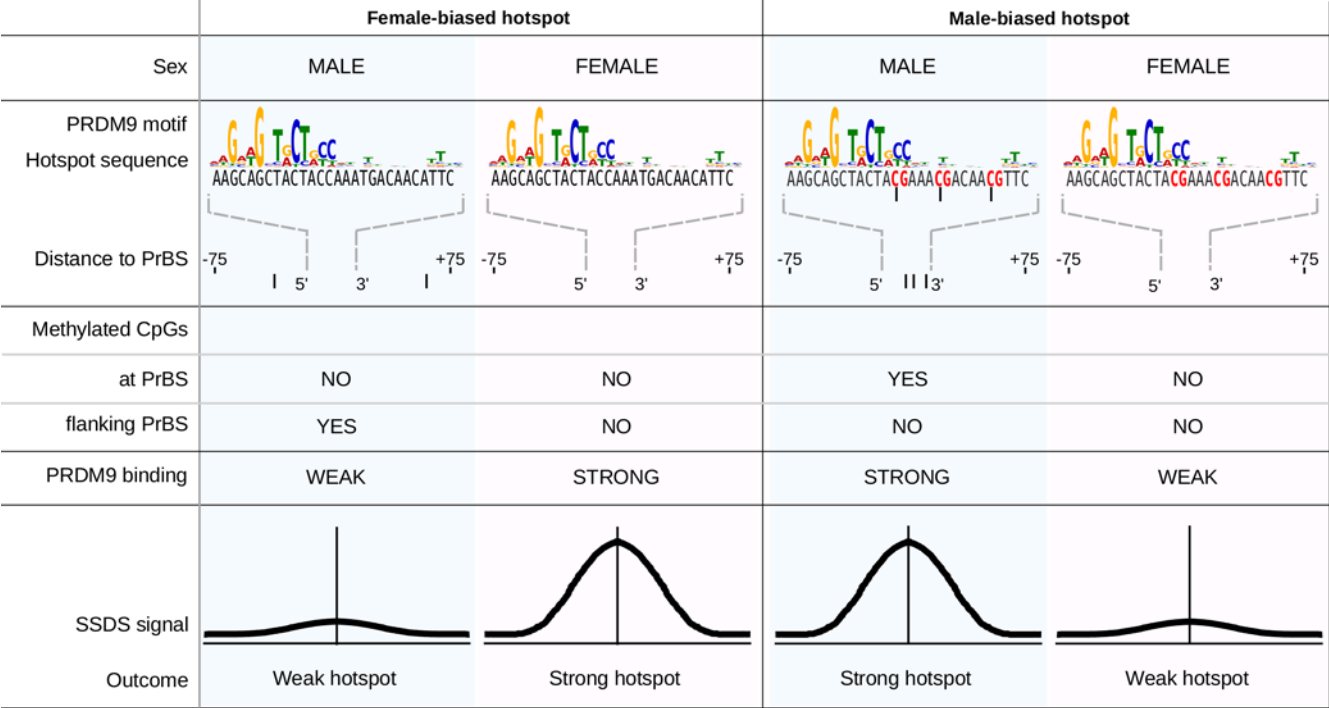
rows depict methylation at individual hotspots. Note that the density of methylation appears higher at unbiased hotspots because rows are more densely spaced.



Extended Data Fig. 8 | See next page for caption.

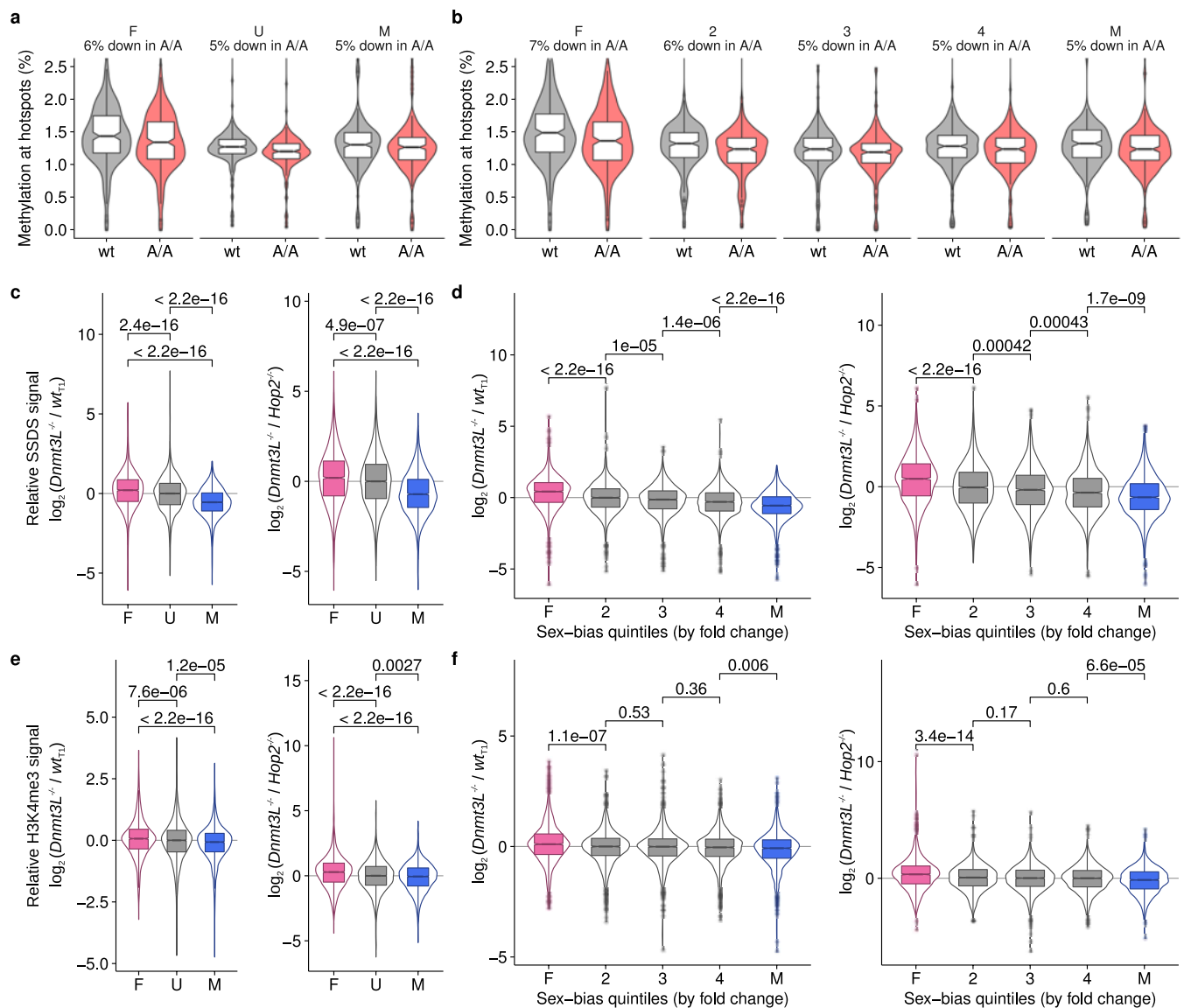
Extended Data Fig. 8 | DNA methylation at PrBSs is present across tissues and absent in the female germ line. The pattern of DNA methylation is very similar across cell types and between the sexes. Hotspots are split by the magnitude of sex bias ($SSDS_{O1}/SSDS_{T1}$) into seven sets. Sets are ranked from most female-biased (pink; left) to most male-biased (blue; right) by fold change. Methylation signal is binarized such that methylation $>0\%$ is considered methylated. Thus, the proportion of all hotspots with methylated cytosine at each position is shown. Variations in the magnitude of the signal may be expected for technical reasons. Plots are anchored by the C57BL/6 PrBS (grey area). Only hotspots with a single PRDM9-binding site are used (see Methods). **a**, Plot of ± 100 bp to show methylation flanking the PrBS for female-biased hotspots.

b, Plot of ± 15 bp to show methylation at the PrBS for male-biased hotspots. Methylation data are from whole-genome bisulfite sequencing (WGBS) in tissue derived from whole testis (Te) in 13 d.p.p. mice⁴⁰, from WGBS after H3K4me3 ChIP-seq in whole adult testis (S/c(Tgt)); Extended Data Fig. 1h), from WGBS in elutriated spermatocytes³⁸ (S/c(Elu)), from WGBS in spermatogonia³⁸ (Gonia), from WGBS in tissue from male liver⁵³ (Liv(M)), from WGBS in tissue from female parous basal differentiated mammary gland cells⁵⁴ (Mamm(F)) and from WGBS in sorted primordial germ cells (PGCs) at 13.5 and 16.5 d.p.c. (PGC13.5(F) and PGC16.5(F), representing earlier and later meiotic prophase I populations, respectively)³⁰. WGBS in female PGCs captures the methylation status of the genome in oocytes during meiosis³⁰.



Extended Data Fig. 9 | Dual role of DNA methylation at hotspots in defining sex biases. DNA methylation has a dual role in modulating sex-biased DSB formation. Left, at female-biased hotspots, DNA methylation in the region flanking the PrBS can suppress PRDM9 binding. Thus, in males, the use of these PrBS is reduced, resulting in a female-biased

hotspot. Methylated CpG dinucleotides (in males) are schematically shown as filled black circles. Right, at male-biased hotspots, DNA methylation at CpGs appears to favour PRDM9 binding and DSB formation. This results in a relatively strong DSB hotspot in males, but a relatively weak hotspot in females, in which DNA methylation at these sites is absent.



Extended Data Fig. 10 | Hotspot strength variation in *Dnmt3l*^{-/-} mice.

a, b, CpG methylation is partly reduced at PRDM9 binding sites in mice lacking functional DNMT3L. We compared WGBS data from *Dnmt3l*^{A/A} (*Dnmt3l*^{D124A/D124A}) (A/A) and matched wild-type mice³³. The ± 100 -bp region around the PRDM9-binding sites was examined. **a–f**, Hotspots were split either by sex-bias (female-biased, F, pink; unbiased, U, grey; male-biased, M, blue) (**a, c, e**) or into quintiles by the fold change between the O1 and T1 SSDS samples (most female-biased (F) on left to most male-biased (M) on right) (**b, d, f**). The percentage decrease in the mean DNA methylation signal in *Dnmt3l*^{A/A} mice for each set is shown. DNA methylation is reduced 5–7%. **c–f**, The usage of sex-biased hotspots is altered in mice in which DNA methylation is reduced (*Dnmt3l*^{-/-}). **c**, Left, the \log_2 fold change between the tags per million normalized signal at hotspots in *Dnmt3l*^{-/-} and wild-type (T1) male mice is shown. Right, to control for spermatocyte population changes resulting from meiotic arrest, we compare to experiments in *Hop2*^{-/-} males instead of wild-type. HOP2 is essential for stabilizing recombination intermediates and mice lacking

functional HOP2 exhibit spermatogenic arrest after DSB formation. Hotspots overlapping gene promoters or default hotspots are excluded as the non-PRDM9 derived H3K4me3 signal would confound these analyses. Furthermore, only hotspots detected in all samples being compared were analysed to remove spurious potential background correlation (**c–f**; $n_{\text{hotspots}} = 9,137$). *P* values for all comparisons are shown (Wilcoxon test). **c**, The SSDS signal at female-biased hotspots is significantly increased in *Dnmt3l*^{-/-} mice compared to male-biased hotspots or unbiased hotspots. The strength of male-biased hotspots is relatively decreased. **d**, This is also seen when we simply split hotspots by fold change. **e**, H3K4me3 at female-biased hotspots is significantly increased in *Dnmt3l*^{-/-} mice compared to male-biased hotspots. H3K4me3 signal at each hotspot was calculated as the sum of overlapping H3K4me3 peak strengths. This is a proxy for DSB hotspot strength, because PRDM9 trimethylates histone H4 lysine 3 before DSB formation. **f**, This is more apparent when we split hotspots into quintiles by sex-bias, probably because H3K4me3 at hotspots is a weak signal.

Crystal structure of the natural anion-conducting channelrhodopsin *GtACR1*

Yoon Seok Kim^{1,9}, Hideaki E. Kato^{2,3,9*}, Keitaro Yamashita⁴, Shota Ito⁵, Keiichi Inoue^{3,5,6}, Charu Ramakrishnan¹, Lief E. Fenno¹, Kathryn E. Evans¹, Joseph M. Paggi^{7,8}, Ron O. Dror^{7,8}, Hideki Kandori^{5,6}, Brian K. Kobilka² & Karl Deisseroth^{1*}

The naturally occurring channelrhodopsin variant anion channelrhodopsin-1 (ACR1), discovered in the cryptophyte algae *Guillardia theta*, exhibits large light-gated anion conductance and high anion selectivity when expressed in heterologous settings, properties that support its use as an optogenetic tool to inhibit neuronal firing with light. However, molecular insight into ACR1 is lacking owing to the absence of structural information underlying light-gated anion conductance. Here we present the crystal structure of *G. theta* ACR1 at 2.9 Å resolution. The structure reveals unusual architectural features that span the extracellular domain, retinal-binding pocket, Schiff-base region, and anion-conduction pathway. Together with electrophysiological and spectroscopic analyses, these findings reveal the fundamental molecular basis of naturally occurring light-gated anion conductance, and provide a framework for designing the next generation of optogenetic tools.

Most organisms depend on light for energy and information. Motile organisms typically capture light using rhodopsin proteins, largely classified into two groups: microbial (type I) and animal (type II)^{1,2}, both exhibiting seven-transmembrane helices and a retinal-based chromophore, but with different effector mechanisms. Animal rhodopsins primarily work as G-protein-coupled receptors that recruit secondary messengers to control effectors such as ion channels that modulate cellular activity, whereas channel and pump microbial rhodopsins can directly provide effector functionality as transmembrane current^{1,2}. Heterologous expression of single-component microbial opsin genes targeted to specific cells of animals defines an experimental approach (optogenetics³) for biology, enabling control of specific cells in behaving organisms with light.

Both channel and pump-encoding opsins are established in optogenetics. Variants of the channel subtype (cation-conducting channelrhodopsins, CCRs) elicit light-triggered cation currents (usually excitatory in neurons). Indeed, light-triggered cation currents are excitatory in the natural host as well; plant behaviours initially observed by botanists more than 150 years ago (movement of single-celled algae excited by light)⁴ were later found to be due to CCRs, with the initially known member of this subclass (*Chlamydomonas reinhardtii* ChR1) identified as a cation channel in 2002⁵. Many CCRs have been discovered or designed^{5–14}, and currently available CCRs offer a palette of diversity in absorption spectrum, photocurrent magnitude, light sensitivity and on/off-kinetics^{12,15}.

The development of inhibitory optogenetics initially lagged, but has made strides in recent years^{3,4,16}. Light-induced neuronal inhibition with microbial opsins was first achieved with inward Cl[−] pumps and outward H⁺ pumps such as *Natronomonas pharaonis* halorhodopsin (NpHR) and archaeorhodopsin-3 (AR3)^{17,18}. Although widely used, these pumps move only one ion per photon (versus hundreds for channels), thereby exhibiting reduced efficacy^{15,16}. In 2014, anion-conducting channelrhodopsins (ACRs) were created^{19,20} on the CCR backbone, guided by structural modelling; subsequently, in 2015,

naturally occurring ACRs were isolated from chlorophyte algae²¹ (*GtACR1* and *GtACR2*). The designed ACRs have been developed further^{22–24}, and additional natural ACRs have been found by genome mining^{25–27}. ACRs can translocate 10⁴–10⁵ ions per second²¹ and can exhibit 10²–10⁴-fold higher light sensitivity than inhibitory pumps^{19–21}. After the first demonstration in 2015 of ACRs as inhibitory optogenetic tools that could successfully modulate animal behaviour (with a designed ACR called iC+²²), both ACR classes have been widely applied in mice, flies and fish^{22,23,28–30}.

Despite progress in ACR-based inhibitory optogenetics, little is known about the structural basis of radically different ion-selectivity involved in anion conduction. Homology models of *GtACR1* were built^{27,31–33} using the structure of the C1C2 CCR³⁴, but precise structural information on ACRs remained completely lacking. A high-resolution crystal structure would be beneficial, not only to enhance fundamental understanding, but also to provide a foundation for expanding the toolbox of optogenetics (as rapidly resulted from the first CCR crystal structure³⁴ in 2012).

Here we obtain and characterize the crystal structure for *GtACR1* at 2.9 Å resolution. This information, together with electrophysiological and spectroscopic analyses, revealed unique natural ACR structure–function relationships that span the extracellular domain, retinal-binding pocket, Schiff base region, and anion-conduction pathway. These features advance our understanding of natural channelrhodopsin biology, and reveal a path for the design and creation of new tools for optogenetics.

Structure determination

To understand the structural basis of light-activated anion conduction, we purified (Extended Data Fig. 1a) and crystallized the best-characterized natural ACR, *GtACR1*. To improve crystallizability, we truncated 13 C-terminal residues; the resulting construct (residues 1–282) showed similar photocurrents to full-length *GtACR1* in human HEK293 cells (Extended Data Fig. 1b) and robust expression in neurons (Extended

¹Department of Bioengineering, Department of Psychiatry and Behavioral Sciences, and Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. ²Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA. ³PRESTO, Japan Science and Technology Agency, Honcho, Kawaguchi, Japan. ⁴RIKEN SPring-8 Center, Hyogo, Japan. ⁵Department of Life Science and Applied Chemistry, Nagoya Institute of Technology, Showa-ku, Nagoya, Japan. ⁶OptoBioTechnology Research Center, Nagoya Institute of Technology, Showa-ku, Nagoya, Japan. ⁷Department of Computer Science, Stanford University, Stanford, CA, USA. ⁸Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. ⁹These authors contributed equally: Yoon Seok Kim, Hideaki E. Kato. *e-mail: hekato@stanford.edu; deisseroth@stanford.edu

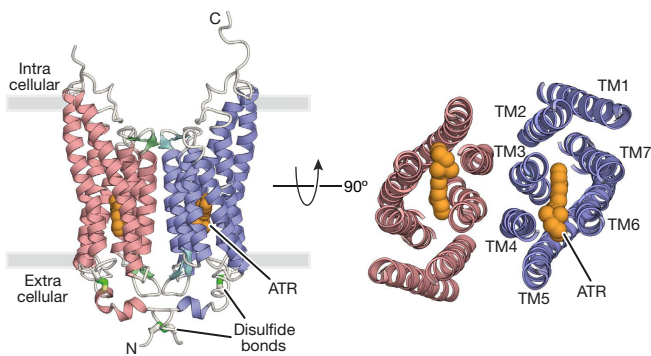


Fig. 1 | Overall structure of *GtACR1*. Crystal structure of the *GtACR1* dimer, viewed parallel to the membrane (left) and from the extracellular side (right). Disulfide bonds are shown using a stick model (green), and ATR (orange) is depicted by a sphere model.

Data Fig. 1c). Crystals were obtained by lipidic cubic phase analysis (Extended Data Fig. 1d); the structure was determined by molecular replacement, using coordinates of C1C2 (Protein Data Bank accession 3UG9)³⁴, and refined to 2.9 Å resolution (Extended Data Fig. 2).

Crystals belonged to the $P2_1$ space group, containing four *GtACR1* protomers (chains A–D) in the asymmetric unit (Extended Data Fig. 1e). Chains A/B and chains C/D were each associated as dimers, with the two dimer molecules arranged anti-parallel. Each protomer showed almost identical conformation except for orientation of certain residues facing the membrane (for example, Trp150, Phe168, Tyr201 and Leu232; Extended Data Fig. 1f), with a notable C-terminal difference. Although the C termini of chains B/C were ordered until Pro273 with similar conformations, those of chains A/D were ordered until Asp278 and Glu280, respectively, and the last 6–8 residues exhibited completely different conformations (Extended Data Fig. 1g). Except for the disordered 3 N-terminal and 2–9 C-terminal residues, *GtACR1* itself (residues 4–278 in chain A, 4–273 in chain B, 4–273 in chain C, and 4–280 in chain D), all-*trans* retinal (ATR), 5 lipids and 4 water molecules were all clearly resolved in the electron density map (Extended Data Fig. 2).

GtACR1 structure and comparison with C1C2 and CrChR2

GtACR1 exhibits a unique N-terminal extracellular domain (residues 4–29), a 7-transmembrane domain (residues 30–249), and a C-terminal region (residues 250–280) (Fig. 1). In comparing *GtACR1* with the CCRs C1C2 (PDB accessions 3UG9 and 4YZI)^{13,34} and CrChR2³⁵, we observed both similarities (despite relatively low sequence identities of 28% and 27%, respectively; Extended Data Fig. 3) and notable distinctions. Although there were aspects of architectural commonality between *GtACR1* and C1C2 dimers and between *GtACR1* and CrChR2 dimers (root mean square deviation (r.m.s.d.) values of 2.10 Å and 1.87 Å respectively over all C_α atoms), and between corresponding monomers (r.m.s.d. values of 1.62 Å and 1.39 Å), many crucial differences with *GtACR1* were apparent (Fig. 2a, b).

First, although transmembrane helix 7 (TM7) of C1C2–CrChR2 protrudes approximately 18 Å from the membrane and its following C-terminal region exhibits a β -sheet (Fig. 2a, b), TM7 of *GtACR1* does not protrude (resembling more pump-type rhodopsins such as bacteriorhodopsin and halorhodopsin) (Extended Data Fig. 4) and its C-terminal region displays a random coil (Fig. 2a, b; Extended Data Fig. 1g); although lacking secondary structure, this region has several hydrogen-bonding interactions with TM5, TM6, intracellular loop 2 (ICL2) and ICL3, and thus could be important in assembly/structural integrity (Extended Data Fig. 5a). To test this, we truncated the corresponding 29 residues from *GtACR1* as crystallized; this almost abolished expression, consistent with the prediction that the C terminus is important for folding and/or stability (Extended Data Fig. 5b).

Second, the N-terminal domain of *GtACR1* has a short helix–loop–helix forming hydrogen-bonding interactions with extracellular

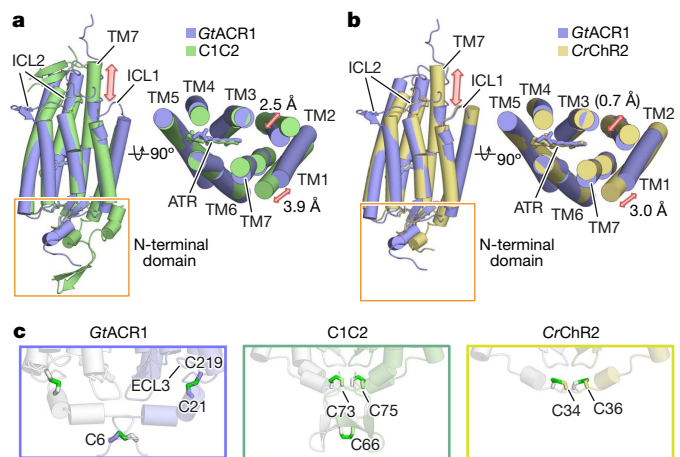


Fig. 2 | Structural comparison of *GtACR1* with C1C2. a, b, Side (left) and extracellular (right) view of *GtACR1* (blue) superimposed onto C1C2 (green) (a), and CrChR2 (yellow) (b). Red arrows mark the differences between the structures. c, Magnified view of N termini of *GtACR1*, C1C2 and CrChR2 as delimited by orange boxes in a and b. Green sticks denote disulfide bonds; note intramolecular disulfide bonds in *GtACR1* (C219-to-C21) compared to the exclusively intermolecular disulfide bonds in C1C2 (at C73, C75, and C66) and CrChR2 (at C34 and C36).

loop 1 (ECL1) (Extended Data Fig. 5c), whereas that of C1C2 has three helices and two β -strands, tethered to ECL1 via both hydrogen bonding and a Zn^{2+} ion¹³. Notably, C1C2, CrChR2 and *GtACR1* all have several (2–3) N-terminal cysteine residues, but with different positions and functions. Cys66, Cys73 and Cys75 of C1C2, and Cys34 and Cys36 of CrChR2 form three and two intermolecular disulfide bridges respectively (Fig. 2c). Previous studies had predicted that the residue corresponding to Cys73 in C1C2 (Cys34 in CrChR2) would be Cys21 in *GtACR1*²⁶ and would form an intermolecular disulfide bridge³². However, the *GtACR1* structure revealed that Cys21 forms not an intermolecular, but instead a new intramolecular, disulfide bridge with Cys219 on ECL3, whereas Cys6 forms an intermolecular disulfide bridge (Figs. 1, 2c). Gel-filtration chromatography and SDS–PAGE (Extended Data Fig. 5d, e) further support the conclusion that Cys21 and Cys219 are more important for folding and expression, and Cys6 for dimerization.

Third, ICL2 of *GtACR1* has a β -sheet that is unique among microbial rhodopsins, extending from the protein core (Fig. 2a, b), in contrast to ICL2 of the C1C2–CrChR2 dimer, which is a random coil close to the protein core involved in dimerization³⁴. Notably, because of these differences in the N terminus and ICL2, the interface area of the *GtACR1* dimer (1,315 Å²) is smaller than that for the C1C2 (2,027 Å²) or CrChR2 (1,688 Å²) dimers (Extended Data Fig. 5f–h). This property is concordant with our finding that loss of the intermolecular disulfide bridge markedly affects *GtACR1* dimerization in SDS–PAGE analysis (Extended Data Fig. 5e), whereas the loss of the disulfide in C1C2–CrChR2 has minimal effect on dimerization^{36–38}.

Finally, we note a feature of overall *GtACR1* structure; the extracellular ends of TM1/TM2 are notably tilted compared to those of CCRs (Fig. 2a, b). These tilts remodel the extracellular vestibule, forming a novel ion-conducting pathway. This unanticipated structural feature appears of substantial importance for understanding the unique ion-conduction properties of *GtACR1* (below).

Retinal-binding pocket

In all rhodopsins, retinal is covalently bound to a TM7 lysine residue, forming the Schiff base. *GtACR1* and C1C2 contain similar configurations of all-*trans*-retinal and 15-*anti*-retinal^{34,39,40} (Fig. 2a). Here we focus on comparison with C1C2, because the 2017 CrChR2 CCR structure was almost identical to the well-studied 2012 C1C2 CCR structure (Extended Data Fig. 4c; r.m.s.d. value of 0.82 Å over all C_α atoms), and was also reported as a mixture of two states D480 and D470

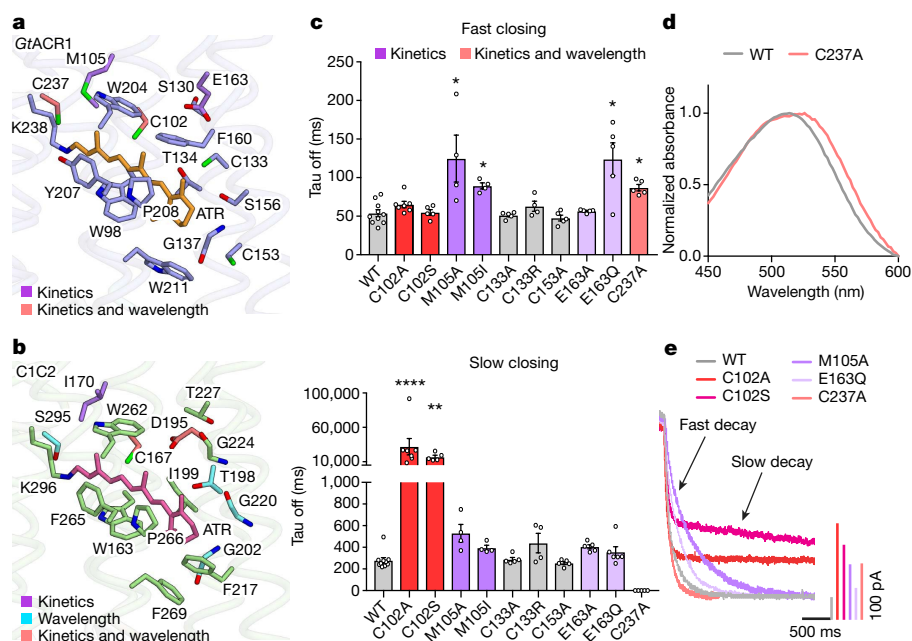


Fig. 3 | RBP of *GtACR1*. **a, b**, RBP of *GtACR1* (**a**) and *C1C2* (**b**). **c**, Effects of mutations (on residues comprising the *GtACR1* RBP) on off-kinetics (top, fast closing; bottom, slow closing). Colour codes summarize the role of each residue in setting kinetics, wavelength or both. Data are mean and s.e.m; $n = 10$ for wild type (WT), 7 for C102A, 4 for

(absorbing light at 480 and 470 nm, respectively)^{35,36,41} making it difficult to compare to *GtACR1*^{35,36}. The *GtACR1* structure reveals that most residues forming the retinal-binding pocket (RBP) are not conserved between *GtACR1* and CCRs (Fig. 3a, b; Extended Data Fig. 3); in *C1C2*, ATR is enclosed by 16 residues (Fig. 3b), but 11 are not conserved in *GtACR1* (Fig. 3a). To analyse the function of these residues, we measured absorption spectra and photocurrents in 10 mutants.

Previous studies reported that *GtACR1* has five spectroscopically distinguishable intermediate states: K, L, M, N and O (with L and M as conducting states), and with opening and closing regulated by two different mechanisms (coupled fast-opening–slow-closing and slow-opening–fast-closing)^{31,33}. Confirming previous measurements, we observed that wild-type *GtACR1* photocurrent peaks at $\lambda_{\max} = 514$ nm with biphasic decay ($\tau_{\text{off1}}: 54 \pm 4.5$ ms; $\tau_{\text{off2}}: 280 \pm 25$ ms), and that mutant *GtACR1*(C102A) shows decelerated τ_{off2} (32 ± 12 s)^{31,33} (Fig. 3c, e). Notably like C102A, C102S also exhibits decelerated τ_{off2} (17 ± 2.5 s). M105A, M105I and E163Q show markedly slowed τ_{off1} (120 ± 30 ms, 90 ± 3.9 ms and 110 ± 22 ms, respectively), suggesting that Met105 and Glu163 are involved in the slow-opening–fast-closing mechanism (Fig. 3c, e).

Notably, studies of *Halobacterium salinarum* bacteriorhodopsin (*HsBR*) predict that mutation of certain residues would affect the energy barrier for the transition from K to L intermediates⁴² (closed to open in *GtACR1*³¹). Thr198, which interacts with the β -ionone ring of ATR in *C1C2*, corresponds to Cys133 in *GtACR1* (Fig. 3a, b). In *HsBR* and *CrChR2*, mutations in residues surrounding the β -ionone affect biophysical properties; for example, M118A in *HsBR* changes the absorption spectrum (λ_{\max} shifting from 551 to 474 nm)⁴³, and T159C in *CrChR2* affects conductance and kinetics⁴⁴. However, the *GtACR1*(C133A) and *GtACR1*(C133R) mutants exhibited only slightly blue-shifted spectra, with kinetics and photocurrents comparable to wild-type levels (Fig. 3c, e; Extended Data Figs. 6–8). Thus, the RBP of the *GtACR1* β -ionone may be unusually robust (which could also depend on additional non-conserved residues around Cys133, such as Thr134 and Phe160; Fig. 3a, b; Extended Data Fig. 3). Another interesting RBP residue is Cys237, which affects key properties including absorption, kinetics and selectivity when mutated to alanine (Fig. 3c–e; Extended Data Figs. 6–8); notably, the mutant exhibits only a single fast

M105A, M105I and C133R, and 5 for the rest. $*P < 0.05$, $**P = 0.0021$, $****P < 0.0001$, Kruskal–Wallis with Dunn's test. **d**, Absorption spectra of wild-type *GtACR1* and the C237A mutant. Spectra were measured in one experiment. **e**, Traces of the wild-type *GtACR1* and four kinetics-shifted mutants. Scale bar denoted by corresponding colour.

component of current decay ($\tau_{\text{off1}}: 87 \pm 4.1$ ms), suggesting involvement of this residue in the slow-closing mechanism (likely along with the Cys102 residue³¹; Fig. 3c–e).

The Schiff–base region

In *C1C2*, two carboxylates (TM3 Glu162, TM7 Asp292) are within 4 Å of the Schiff-base nitrogen, which forms a direct hydrogen bond with Asp292 (Fig. 4a). However, in *GtACR1*, the TM3 residue is Ser97, and the TM7 Asp234 has a conformation quite different from Asp292 of *C1C2*, possibly owing to local interactions with Tyr72 and Tyr207. Notably, the overall architecture of the Schiff-base region in *GtACR1* is more similar to halorhodopsins (Fig. 4a). However, in *GtACR1*, there is no clear electron density, suggesting water or Cl^- within hydrogen-bonding distance of the Schiff-base (Supplementary Discussion), and presumably the protonated Schiff base forms at least a weak hydrogen bond with Asp234 (Fig. 4a). Therefore, we undertook structure-guided functional characterization of Tyr72, Tyr207 and Asp234.

First, we analysed protonation of Asp-234 using ultraviolet-visible (UV-vis) and low-temperature Fourier-transform infrared (FTIR) spectroscopy. Both assays strongly suggested the protonation of Asp234 in the dark, for the following reasons: first, wild-type and D234N mutants showed almost identical UV-vis absorption spectra (Fig. 4b; Extended Data Fig. 9a); and second, the light-induced difference-FTIR spectra at 77 K showed that a peak pair at $1,740(-)/1,732(+)$ cm^{-1} in the wild type, assigned to C=O vibration of a protonated carboxylate^{39,45}, disappears in the D234N mutant (Fig. 4c; Extended Data Fig. 9b). Because the wild-type λ_{\max} of the UV-vis spectra and intensity of the FTIR peak-pair remain unchanged from pH 5–9 (Extended Data Fig. 9c), Asp234 is therefore presumed to be protonated over a wide pH range, concordant with previous Raman spectroscopy³⁹.

However, surprisingly, electrophysiology revealed that D234N nearly abolishes the photocurrent (Fig. 4d). Generally, the effects of aspartate-to-asparagine mutation are small when aspartate is protonated, but in the uniquely configured *GtACR1* Schiff-base environment involving close apposition of Asp234, the small difference between aspartate-hydroxyl and asparagine-amino could rearrange the hydrogen-bond network around the Schiff base and thus disturb light-induced conformational changes. This concept is supported by difference FTIR spectra

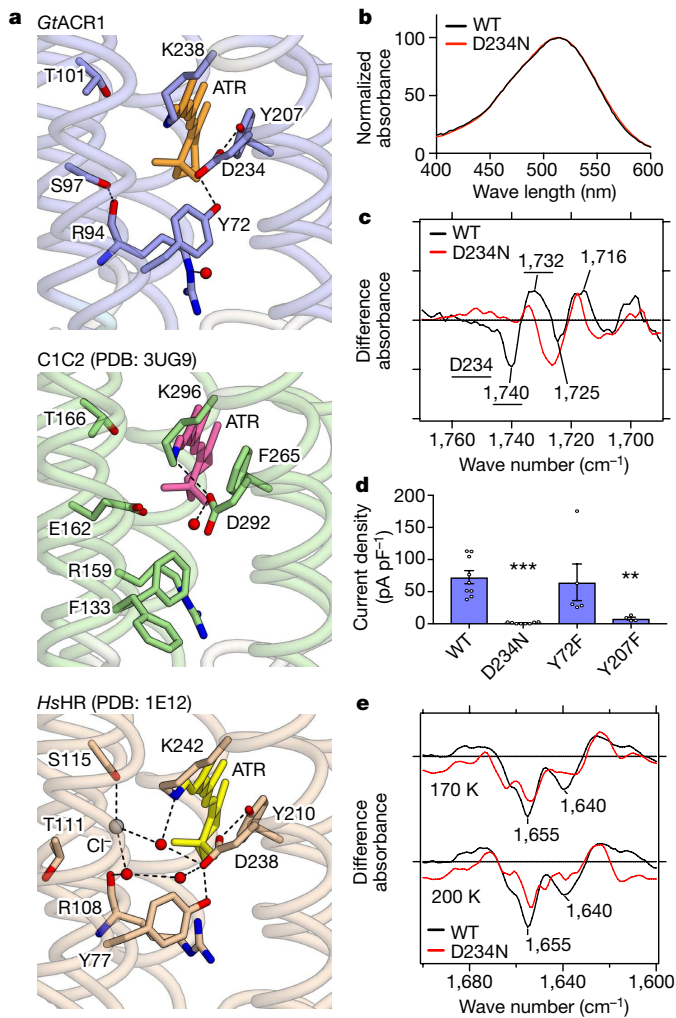


Fig. 4 | The protonated Schiff base region of *GtACR1* and its counterions. **a**, Structures of the Schiff base in *GtACR1* (top), C1C2 (middle) and *HsHR* (bottom). Red spheres and dashed lines represent water molecules and hydrogen bonds, respectively; in *GtACR1*, D234 forms hydrogen bonds with the protonated Schiff base, Y72 and Y207, more similarly to *HsHR* than C1C2. **b**, Similar absorption spectra of wild-type *GtACR1* and the D234N mutant, suggesting D234 protonation in the dark (see also Extended Data Fig. 9a). **c**, Light-induced difference FTIR spectra at 77 K. Note disappearance of the 1,740(–)/1,732(+) cm^{-1} peak pair (assigned to C=O vibration of a protonated carboxylate^{39,45}) in D234N. Findings in **b** and **c** hold from at least pH 5–9 (Extended Data Fig. 9c). **d**, Current densities of wild-type *GtACR1* and three mutants. Note D234N abolishes the photocurrent (surprising if protonated in the dark), and Y207 (but not Y72) is essential (consistent with the importance of the local hydrogen-bonded network). Data are mean and s.e.m.; $n = 9$ for WT, 8 for D234N, 5 for Y72F and 4 for Y207F. $^{**}P = 0.01$, $^{***}P = 0.0006$, one-way ANOVA followed by Dunnett's test. **e**, Light-induced difference FTIR spectra of wild type and D234N at 170 K and 200 K. Decreased intensity of negative bands at 1,640 and 1,655 cm^{-1} reveals smaller conformational change of transmembrane helices in D234N. All spectroscopy experiments were performed once.

in the amide-I region at 170 K and 200 K (Fig. 4e): the intensity of negative bands at 1,640 and 1,655 cm^{-1} decreases in D234N, revealing that the conformational change of transmembrane helices in D234N is significantly smaller than in the wild type. Just as with D234N, the nearby Y207F mutation also causes loss-of-function (Fig. 4d). Considering that Phe207 naturally occurs in fully functional C1C2 and even in other natural ACRs including *GtACR2* and the ZipACR variant with divergent sequences²⁷ (Extended Data Fig. 3), the precisely arranged hydrogen-bond network of the Schiff-base region thus appears essential for channel activity.

Ion conducting pathway and constrictions

To identify the ion-conduction pathway, we calculated the full electrostatic surface potential of *GtACR1* compared to C1C2. C1C2 has a cation-conducting pore pathway formed by TM1, TM2, TM3 and TM7, and *GtACR1* has a pore pathway at approximately the same position (Fig. 5a, b) with three marked differences. First, in a pattern opposite to that of C1C2, the surface around the pore of *GtACR1* is electropositive, suitable for cation exclusion and thus anion selectivity⁴⁶ (Fig. 5a; Extended Data Fig. 10a); by contrast, C1C2 has 7 carboxylates along the ion-conducting pathway (Glu121, Glu122, Glu129, Glu136, Glu140, Glu162 and Asp292) and 14 carboxylates on intracellular/extracellular surfaces (Fig. 5b; Extended Data Fig. 10b, d), all contributing to electronegative surfaces in and around the pore suitable for anion exclusion/cation selectivity (Fig. 5b). In *GtACR1*, Glu122, Glu136 and Glu162 are replaced by Ala61, Ala75 and Ser97, respectively (Fig. 5a) and Glu140 is also not conserved (Extended Data Fig. 3). Also, as shown by previous and present FTIR, residues corresponding to Glu129/Asp292 (Glu68/Asp234) are neutralized⁴⁵ (Fig. 4b, c). Finally, 12 protein-surface residues are replaced with arginine or lysine; the consistency of this pattern suggests these residues (and Arg94/Lys238) contribute to a suitable electrostatic environment for cation exclusion/anion conduction in *GtACR1*, confirming earlier predictions⁴⁶ (Extended Data Figs. 3, 10a, c).

Second, extracellular vestibules of *GtACR1* differ markedly from C1C2. C1C2 has two extracellular vestibules (EV1 and EV2) but only EV2 is connected to the ion-conducting pathway; EV1 is occluded by hydrogen bonding among Gln95, Glu136 and Glu140 (extracellular constriction site 1, ECS1) (Fig. 5b). However, Glu136 and Glu140 are not conserved in *GtACR1*, and the extracellular-side TM1 and TM2 are markedly tilted, as described above (Figs. 2c, 5a; Extended Data Fig. 3). Thus, pore size becomes much larger, and EV1 becomes connected to the *GtACR1* pore-pathway. Furthermore, in contrast to EV1, EV2 of *GtACR1* is disconnected because of interactions among Tyr81, Arg94 and Glu223 (extracellular constriction site 2, ECS2) (Figs. 5a, 6a), indicating that EV1 serves as the primary anion-entry pathway in *GtACR1*. Third, the anion-conducting pathway of *GtACR1* is opened not only towards the extracellular side but also intracellularly. In C1C2, although the cation-conducting pathway is opened towards the extracellular side, the cytoplasmic side is occluded by intracellular (ICS) and central (CCS) constriction sites³⁴. However, in *GtACR1*, residues forming the ICS in C1C2, including Tyr109, Glu122, His173 and Arg307, are replaced by Met 50, Ala61, Leu108 and Thr249, respectively, and the intracellular vestibule extends to the CCS (Figs. 5, 6b, c).

The channel is thus maintained in a closed state only by the CCS (Figs. 5, 6c). In C1C2, the CCS is formed by Ser102, Glu129 and Asn297. These three residues are conserved in *GtACR1* (Gln46, Glu68 and Asn239) and its Gln46 on TM1 forms an additional hydrogen bond with Asn239, thereby further stabilizing the CCS. To test the function of these residues, we prepared 10 mutants of Gln46, Glu68 and Asn239, and measured activity by patch-clamp analysis. All Glu68 and Asn239 mutants exhibited smaller photocurrents, and Q46A showed comparable photocurrents but depolarized reversal-potential (Fig. 6d, e). Thus, all three CCS residues are important for anion-channel function, but with different roles: Glu68 and Asn239 for conductance, and Gln46 for selectivity.

Discussion

This high-resolution view into the inner workings of *GtACR1* reveals that CCRs and natural ACRs share certain overall features, but also exhibit highly informative differences (especially in the architecture of the *GtACR1* anion-conducting pathway, with exchange of one extracellular vestibule for another). The *GtACR1* closed-state pore is also remarkable, almost entirely open with the exception of a single central constriction formed by Gln46, Asn239, Ser43 and Glu68; anions can be released intracellularly via the open conduction pore formed by Ala61, Leu108 and Thr249 (Figs. 5a, 6b). Thus, these data provide the first,

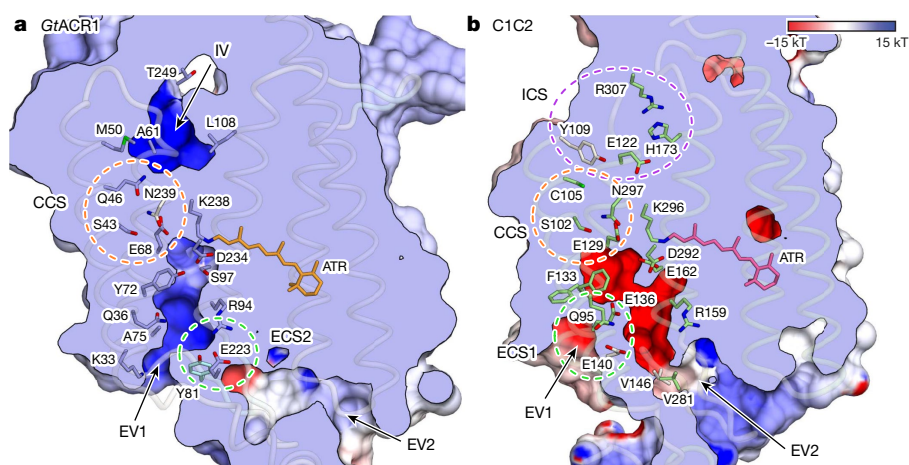


Fig. 5 | Ion-conducting pathways of *GtACR1* and *C1C2*. **a, b**, Ion-conducting pathways of *GtACR1* (**a**) and *C1C2* (**b**). The surface is coloured by the electrostatic potential calculated using PDB accession 2PQR⁵¹ for both *GtACR1* and *C1C2*. Green, purple and orange-dashed

circles represent the extracellular constriction site (ECS), intracellular constriction site (ICS) and central constriction site (CCS), respectively. IV, intracellular vestibule.

to our knowledge, crystal structure of any channelrhodopsin revealing an open intracellular pore pathway.

Integration of structural, electrophysiological and spectroscopic analyses uncovered unique features of the Schiff base relevant to ChR (and halorhodopsin and bacteriorhodopsin) evolution. As in *HsHR*, a TM7 aspartate is coordinated by two tyrosine residues in *GtACR1*, and the TM3 glutamate in the CCR *C1C2* is instead represented in both *GtACR1* and *HsHR* by a neutral hydrophilic residue (Fig. 4a). Furthermore, a TM2 tyrosine (Tyr72, uniformly conserved among pump-type halorhodopsins and bacteriorhodopsins) is present in *GtACR1* (and is almost 100% conserved among natural ACRs; Extended Data Fig. 3)²⁷ but is dispensable for function; Y72F changes neither conductance (Fig. 4d) nor kinetics of the M-intermediate rise or decay (Extended Data Fig. 6d), characterized by fast or slow kinetics, respectively. This differs from bacteriorhodopsin, in which Y57F accelerates formation of the M-intermediate⁴⁷. Because CCRs have replaced this residue (Extended Data Fig. 3), an evolutionary model is suggested in which natural ACRs such as *GtACR1* evolved from light-driven Cl[−] pumps, and CCRs subsequently arose from natural ACRs via surface electrostatic remodelling^{4,46}.

Further insight into the mechanism and development of anion conduction arises from the consideration of another unusual feature of the Schiff-base region: charge distribution. In the dark, the Schiff-base nitrogen is protonated and therefore requires a mechanism to stabilize the positive charge. In *GtACR1*, Glu68 and Asp234 provide the only carboxylates within 6 Å of this Schiff-base nitrogen (approximately 5.4 Å and 3.5 Å, respectively), but FTIR analyses indicate that both are

also protonated in the dark⁴⁵ (Fig. 4d, e). Cl[−] does not have a charge-stabilization role either, as Cl[−] is not bound to the Schiff-base region in *GtACR1*³³ (Fig. 4a; unlike in *HsHR*⁴⁸). One possible explanation is that strongly polarized water could bind to the Schiff base (behaving as a hydroxyl ion, as proposed in *HsBR* and mutants^{2,49,50}; Supplementary Discussion), and another possibility is that the partial-negative charge of the nearby Asp234 carbonyl is sufficient to weakly stabilize the positively charged Schiff base. As a result, the net charge in the Schiff-base region may represent the achievement of perhaps the most challenging evolutionary step in the adaptation to facilitate anion conduction (alongside the acquisition of positive surface electrostatic potential throughout the pore and vestibules; Fig. 4): namely, partial local positivity despite the obligate negative nature of the Schiff base counterion.

To advance our understanding of the molecular mechanism of light-gated anion conduction, additional studies (including structural resolution of natural or designed ACRs in fully open or intermediate states) will be required. This initial high-resolution structural information provides a framework for the further development of ACR-based optogenetic tools—for example, the creation of kinetic, spectral and selectivity variants that maintain the advantages of the *GtACR1* backbone including strong photocurrents, just as the initial CCR structure³⁴ allowed the development of new classes of optogenetic functionality⁴. Further insights into the evolutionary and functional relationships among different channelrhodopsin family members will continue to arise from the solution of structures that correspond to kinetic, spectral and selectivity variants, advancing basic understanding of this remarkable class of natural protein.

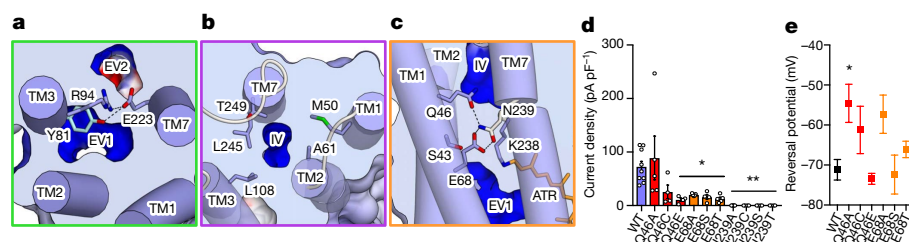


Fig. 6 | Constriction sites of *GtACR1*. **a**, The ECS separating EV1 and EV2. Hydrogen bonds are shown as dashed lines. **b**, Initial glimpse of a patent intracellular conduction pathway for a light-activated channel; architecture of the *GtACR1* intracellular ion exit pore leading to the intracellular vestibule (IV). **c**, The CCS architecture: sole constriction site in the pore, which separates the extracellular and intracellular vestibules. **d**, Current densities of mutants in residues comprising the CCS. Note the importance of residues E68 and N239 for photocurrents. Data are mean

and s.e.m. $n = 9$ for WT, 5 for Q46A, E68A, E68T and E239A, and 4 for the rest. $*P < 0.05$, $**P < 0.01$, one-way ANOVA followed by Dunnett's test. **e**, Comparison of reversal potentials. Note the signature of increased cation flux (depolarized reversal potential), consistent with disrupted pore selectivity. Data are mean and s.e.m. $n = 10$ for WT, 6 for Q46A and Q46C, 5 for E68A and 4 for the rest. $*P = 0.014$, one-way ANOVA followed by Dunnett's test.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0511-6>.

Received: 23 February 2018; Accepted: 13 August 2018;

Published online: 29 August 2018

- Zhang, F. et al. The microbial opsin family of optogenetic tools. *Cell* **147**, 1446–1457 (2011).
- Ernst, O. P. et al. Microbial and animal rhodopsins: structures, functions, and molecular mechanisms. *Chem. Rev.* **114**, 126–163 (2014).
- Deisseroth, K. Optogenetics: 10 years of microbial opsins in neuroscience. *Nat. Neurosci.* **18**, 1213–1225 (2015).
- Deisseroth, K. & Hegemann, P. The form and function of channelrhodopsin. *Science* **357**, eaan5544 (2017).
- Nagel, G. et al. Channelrhodopsin-1: a light-gated proton channel in green algae. *Science* **296**, 2395–2398 (2002).
- Nagel, G. et al. Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proc. Natl Acad. Sci. USA* **100**, 13940–13945 (2003).
- Zhang, F. et al. Red-shifted optogenetic excitation: a tool for fast neural control derived from *Volvox carter*. *Nat. Neurosci.* **11**, 631–633 (2008).
- Berndt, A., Yizhar, O., Gunaydin, L. A., Hegemann, P. & Deisseroth, K. Bi-stable neural state switches. *Nat. Neurosci.* **12**, 229–234 (2009).
- Gunaydin, L. A. et al. Ultrafast optogenetic control. *Nat. Neurosci.* **13**, 387–392 (2010).
- Yizhar, O. et al. Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* **477**, 171–178 (2011).
- Lin, J. Y., Knutsen, P. M., Muller, A., Kleinfeld, D. & Tsien, R. Y. ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nat. Neurosci.* **16**, 1499–1508 (2013).
- Klapoetke, N. C. et al. Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).
- Kato, H. E. et al. Atomistic design of microbial opsin-based blue-shifted optogenetics tools. *Nat. Commun.* **6**, 7177 (2015).
- Rajasekharan, P. et al. Projections from neocortex mediate top-down control of memory retrieval. *Nature* **526**, 653–659 (2015).
- Mattis, J. et al. Principles for applying optogenetic tools derived from direct comparative analysis of microbial opsins. *Nat. Methods* **9**, 159–172 (2011).
- Wiegert, J. S., Mahn, M., Prigge, M., Printz, Y. & Yizhar, O. Silencing neurons: tools, applications, and experimental constraints. *Neuron* **95**, 504–529 (2017).
- Zhang, F. et al. Multimodal fast optical interrogation of neural circuitry. *Nature* **446**, 633–639 (2007).
- Chow, B. Y. et al. High-performance genetically targetable optical neural silencing by light-driven proton pumps. *Nature* **463**, 98–102 (2010).
- Berndt, A., Lee, S. Y., Ramakrishnan, C. & Deisseroth, K. Structure-guided transformation of channelrhodopsin into a light-activated chloride channel. *Science* **344**, 420–424 (2014).
- Wietek, J. et al. Conversion of channelrhodopsin into a light-gated chloride channel. *Science* **344**, 409–412 (2014).
- Govorunova, E. G., Sineshchekov, O. A., Janz, R., Liu, X. & Spudich, J. L. Natural light-gated anion channels: a family of microbial rhodopsins for advanced optogenetics. *Science* **349**, 647–650 (2015).
- Berndt, A. et al. Structural foundations of optogenetics: determinants of channelrhodopsin ion selectivity. *Proc. Natl Acad. Sci. USA* **113**, 822–829 (2016).
- Wietek, J. et al. An improved chloride-conducting channelrhodopsin for light-induced inhibition of neuronal activity *in vivo*. *Sci. Rep.* **5**, 14807 (2015).
- Wietek, J. et al. Anion-conducting channelrhodopsins with tuned spectra and modified kinetics engineered for optogenetic manipulation of behavior. *Sci. Rep.* **7**, 14957 (2017).
- Govorunova, E. G., Sineshchekov, O. A. & Spudich, J. L. *Proteomonas sulcata* ACR1: a fast anion channelrhodopsin. *Photochem. Photobiol.* **92**, 257–263 (2016).
- Wietek, J., Broser, M., Krause, B. S. & Hegemann, P. Identification of a natural green light absorbing chloride conducting channelrhodopsin from *Proteomonas sulcata*. *J. Biol. Chem.* **291**, 4121–4127 (2016).
- Govorunova, E. G. et al. The expanding family of natural anion channelrhodopsins reveals large variations in kinetics, conductance, and spectral sensitivity. *Sci. Rep.* **7**, 43358 (2017).
- Iyer, S. M. et al. Optogenetic and chemogenetic strategies for sustained inhibition of pain. *Sci. Rep.* **6**, 30570 (2016).
- Zhao, Z. et al. A central catecholaminergic circuit controls blood glucose levels during stress. *Neuron* **95**, 138–152 (2017).
- Mohammad, F. et al. Optogenetic inhibition of behavior with anion channelrhodopsins. *Nat. Methods* **14**, 271–274 (2017).
- Sineshchekov, O. A., Govorunova, E. G., Li, H. & Spudich, J. L. Gating mechanisms of a natural anion channelrhodopsin. *Proc. Natl Acad. Sci. USA* **112**, 14236–14241 (2015).
- Li, H., Sineshchekov, O. A., Wu, G. & Spudich, J. L. *In vitro* activity of a purified natural anion channelrhodopsin. *J. Biol. Chem.* **291**, 25319–25325 (2016).
- Sineshchekov, O. A., Li, H., Govorunova, E. G. & Spudich, J. L. Photochemical reaction cycle transitions during anion channelrhodopsin gating. *Proc. Natl Acad. Sci. USA* **113**, E1993–E2000 (2016).
- Kato, H. E. et al. Crystal structure of the channelrhodopsin light-gated cation channel. *Nature* **482**, 369–374 (2012).
- Volkov, O. et al. Structural insights into ion conduction by channelrhodopsin 2. *Science* **358**, eaan8862 (2017).
- Krause, N., Engelhard, C., Heberle, J., Schlesinger, R. & Bittl, R. Structural differences between the closed and open states of channelrhodopsin-2 as observed by EPR spectroscopy. *FEBS Lett.* **587**, 3309–3313 (2013).
- Sattig, T., Rickert, C., Bamberg, E., Steinhoff, H. J. & Bamann, C. Light-induced movement of the transmembrane helix B in channelrhodopsin-2. *Angew. Chem. Int. Edn Engl.* **52**, 9705–9708 (2013).
- Pescitelli, G. et al. Exciton circular dichroism in channelrhodopsin. *J. Phys. Chem. B* **118**, 11873–11885 (2014).
- Yi, A., Mamaeva, N., Li, H., Spudich, J. L. & Rothschild, K. J. Resonance raman study of an anion channelrhodopsin: effects of mutations near the retinylidene Schiff base. *Biochemistry* **55**, 2371–2380 (2016).
- Hontani, Y. et al. Reaction dynamics of the chimeric channelrhodopsin C1C2. *Sci. Rep.* **7**, 7217 (2017).
- Bruun, S. et al. Light–dark adaptation of channelrhodopsin involves photoconversion between the all-trans and 13-cis retinal isomers. *Biochemistry* **54**, 5389–5400 (2015).
- Maeda, A., Tomson, F. L., Gennis, R. B., Balashov, S. P. & Ebrey, T. G. Water molecule rearrangements around Leu93 and Trp182 in the formation of the L intermediate in bacteriorhodopsin's photocycle. *Biochemistry* **42**, 2535–2541 (2003).
- Greenhalgh, D. A., Farrens, D. L., Subramaniam, S. & Khorana, H. G. Hydrophobic amino acids in the retinal-binding pocket of bacteriorhodopsin. *J. Biol. Chem.* **268**, 20305–20311 (1993).
- Berndt, A. et al. High-efficiency channelrhodopsins for fast neuronal stimulation at low light levels. *Proc. Natl Acad. Sci. USA* **108**, 7595–7600 (2011).
- Yi, A. et al. Structural changes in an anion channelrhodopsin: formation of the K and L intermediates at 80 K. *Biochemistry* **56**, 2197–2208 (2017).
- Berndt, A. & Deisseroth, K. Expanding the optogenetics toolkit. *Science* **349**, 590–591 (2015).
- Govindjee, R. et al. Effects of substitution of tyrosine 57 with asparagine and phenylalanine on the properties of bacteriorhodopsin. *Biochemistry* **34**, 4828–4838 (1995).
- Kolbe, M., Besir, H., Essen, L. O. & Oesterhelt, D. Structure of the light-driven chloride pump halorhodopsin at 1.8 Å resolution. *Science* **288**, 1390–1396 (2000).
- Betancourt, F. M. & Glaeser, R. M. Chemical and physical evidence for multiple functional steps comprising the M state of the bacteriorhodopsin photocycle. *Biochim. Biophys. Acta* **1460**, 106–118 (2000).
- Facciotti, M. T., Rouhani, S. & Glaeser, R. M. Crystal structures of bR(D85S) favor a model of bacteriorhodopsin as a hydroxyl-ion pump. *FEBS Lett.* **564**, 301–306 (2004).
- Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32**, W665–W667 (2004).

Acknowledgements We thank C. Lee, M. Lo, K. Geiselhart and M. Lima for technical support; K. K. Kumar, N. R. Latorraca, M. Inoue and K. Katayama for critical comments; and the APS beamline staff at 23ID-B and 23ID-D for assistance in data collection. We acknowledge support by the Stanford Bio-X and the Kwanjeong Foundation (Y.S.K.), JST PRESTO (JPMJPR1782 to H.E.K., JPMJPR15P2 to K.I.), the US Department of Energy, Scientific Discovery through Advanced Computing (SciDAC) program (R.O.D.), MEXT (17H03007 to K.I.), 25104009/15H02391 for H.K.), J.S.T. CREST (JPMJCR1753, H.K.) and Mathers Charitable Foundation (B.K.K.). K.D. was supported by a grant for channelrhodopsin crystal structure determination from the NIMH (R01MH075957 to K.D.).

Reviewer information Nature thanks P. Scheerer, L. Tian and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.S.K. and H.E.K. contributed equally and either has the right to list himself first in bibliographic documents. Y.S.K. and H.E.K. expressed, purified and crystallized GtACR1, harvested crystals, and collected diffraction data. H.E.K. and K.Y. processed the diffraction data and solved the structure. Y.S.K. and L.E.F. performed electrophysiology. Y.S.K. measured UV-vis spectra. S.I. performed FTIR experiments under the guidance of K.I. and H.K. J.M.P. and R.O.D. provided input on structural considerations. C.R. and K.E.E. performed cell cultures and molecular cloning for electrophysiology. K.D. initiated and supervised this ChR structure/function project; Y.S.K., H.E.K., B.K.K. and K.D. planned and guided the work, and interpreted the data. Y.S.K., H.E.K. and K.D. prepared the manuscript and wrote the paper with input from all the authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0511-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0511-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to H.E.K. or K.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample sizes were determined based on previous literature and best practices in the field; no statistical methods were used to predetermine sample size. No experiments in animals were conducted in this paper and hence experiments were not randomized or blinded.

Cloning, protein expression and purification. The crystallization construct of *GtACR1* was generated with several features to enhance protein purification and crystallogenesis. The flexible 13 amino acids at the C terminus were truncated after Gly282. A Flag tag followed by the 3C protease cleavage site was added to the N terminus and an enhanced GFP (eGFP) with a His₁₀ tag and the 3C site was attached to the truncated C terminus via the 3C cleavage site. The finalized *GtACR1* crystallization construct was expressed in Sf9 cells using the BestBac (Expression Systems) baculovirus system. Cell cultures were grown to a density of 4×10^6 cells ml⁻¹, infected with *GtACR1* baculovirus, and shaken at 27°C for 18 h. Then, 20 µM *all-trans* retinal (ATR) (Sigma) was supplemented to the culture and incubation continued for 42 more hours, and cell pellets were collected and stored at -80°C. To purify *GtACR1*, the pellets were lysed with a hypotonic lysis buffer (20 mM HEPES pH 7.5, 1 mM EDTA and protease inhibitors). The cell debris was then homogenized with a glass douncer in a solubilization buffer (1% *n*-dodecyl-β-D-maltopyranoside (DDM), 0.06% cholesteryl hemisuccinate tris salt (CHS), 20 mM HEPES pH 7.5, 500 mM NaCl, 20% glycerol, 10 mM imidazole and protease inhibitors) and solubilized for 2 h in 4°C. The insoluble cell debris was removed by centrifugation (38,000g, 25 min), and the supernatant was mixed with the Ni-NTA agarose resin (Qiagen) for 2 h in 4°C. The Ni-NTA resin was collected into a glass chromatography column, washed with 20 column volumes of a wash buffer (0.05% DDM, 0.01% CHS, 20 mM HEPES pH 7.5, 500 mM NaCl, 20% glycerol and 20 mM imidazole) and was eluted in a wash buffer supplemented with 250 mM imidazole. The Ni-NTA eluent was then supplemented with 2 mM CaCl₂ and was loaded over anti-Flag M1 resin over 1 h. The protein was then washed with a Flag wash buffer (0.05% DDM, 0.01% CHS, 20 mM HEPES pH 7.5, 300 mM NaCl, 5% glycerol and 2 mM CaCl₂) and eluted with a Flag elution buffer (0.05% DDM, 0.01% CHS, 20 mM HEPES pH 7.5, 300 mM NaCl, 5% glycerol, 0.2 mg ml⁻¹ Flag peptide and 3 mM EDTA). After the cleavage of the Flag tag and eGFP-His₁₀ by His-tagged 3C protease, the sample was reloaded onto the Ni-NTA column to capture the cleaved eGFP-His₁₀. The flow-through containing *GtACR1* was collected, concentrated and purified through gel-filtration chromatography in a final buffer (100 mM NaCl, 20 mM HEPES pH 7.5, 0.05% DDM and 0.01% CHS). Peak fractions were pooled and concentrated to 30 mg ml⁻¹ (Extended Data Fig. 1b).

Crystallization. Purified *GtACR1* protein was crystallized using the lipidic cubic phase (LCP) method as described previously³⁴. Protein was mixed with monopalmitolein (Nu-chek) at a weight ratio of 1:1 (protein:lipid) using a coupled syringe mixing device. Then, 20–25 nl protein–LCP mixture drops were accurately dispensed on a 96-well sandwich plate and overlaid by 500 nl of precipitant solution by the Gryphon LCP robot (Art Robbins Instruments). Initial crystals were obtained in 10% (w/v) polypropylene glycol P 400 (PPG P400), 100 mM MES pH 6.0 and 100 mM potassium formate; the best crystals were obtained in 10–12% (w/v) polypropylene glycol P 400 (PPG P400), 100 mM MES pH 6.0, 100 mM potassium formate and 1–3% 1-butanol. Crystals were harvested using micromeshes (MiTeGen), and were flash-cooled in liquid nitrogen without any additional cryoprotection.

Data collection and structure determination. X-ray diffraction data were collected at Advanced Photon Source GM/CA-CAT beamline 23ID-B and 23ID-D using a micro beam size of 10×10 µm², at a wavelength of 1.033 Å. Small wedge data, each consisting of 5–20°, were collected from single crystals, and 131 collected datasets were processed automatically using KAMO⁵². Each dataset was indexed and integrated using XDS⁵³, and classified using the correlation coefficients between data sets. Eighty datasets in the best cluster were scaled and merged using XSCALE. The structure was determined by molecular replacement with the program MoRDa (Vagin and Lebedev; <http://www.biomexsolutions.co.uk/morda>), using the cation channelrhodopsin C1C2 (PDB accession 3UG9) and G11A mutant of SARS-CoV 3C-like protease (PDB accession 2PWX) as the search models. However, the 2PWX model was not fitted to electron density at all and removed. The resultant structure was iteratively refined using Refmac⁵⁴, Phenix⁵⁵ and MR-rosetta⁵⁶, and manually rebuilt in Coot⁵⁷. The final model contained 95.7, 4.1 and 0.3% in the favoured, allowed and outlier regions of the Ramachandran plot, respectively. Final refinement statistics are summarized in Extended Data Fig. 1. All molecular graphics figures were prepared with Cuemol (Ishitani; <http://www.cuemol.org>).

Electrophysiology. HEK293 cells (Thermo Fisher, authenticated by the vendor, not tested for mycoplasma contamination) were plated on poly-D-lysine coated glass coverslips (Fisher) at 10% confluency, and were transfected with 0.5 µg of a plasmid and 1 µl lipofectamine 2000 (ThermoFisher Scientific) per well. After 24–48 h of

transfection, cells were placed in an extracellular tyrode medium (150 mM NaCl, 4 mM KCl, 2 mM CaCl₂, 2 mM MgCl₂, 10 mM HEPES pH 7.4 and 10 mM glucose). A borosilicate patch pipette (Harvard Apparatus) with resistance of 3–6 MΩ was filled with intracellular medium (140 mM potassium-gluconate, 10 mM EGTA, 2 mM MgCl₂ and 10 mM HEPES pH 7.2). The photocurrent and kinetic measurements were performed in voltage-clamp mode at membrane potential of -70 mV and -10 mV, respectively. Light was delivered with the Spectra X Light engine (Lumencor) connected to the fluorescence port of a Leica DM LFSA microscope, and a 513/15 filter was used for green light generation. To determine channel kinetics and photocurrent amplitudes, traces were first smoothed using a lowpass Gaussian filter with a -3 dB cutoff for signal attenuation and noise reduction at 1,000 Hz and then analysed in Clampfit software (Axon Instruments). Liquid junction potentials were corrected using the Clampex built-in liquid junction potential calculator as previously described²². Current density was calculated by dividing peak photocurrent amplitude by cell's membrane capacitance, which was calculated from the Clampex built-in membrane test. Statistical analysis was performed with *t*-test or one-way ANOVA, and the Kruskal–Wallis test for non-parametric data, using Prism 7 (GraphPad) software.

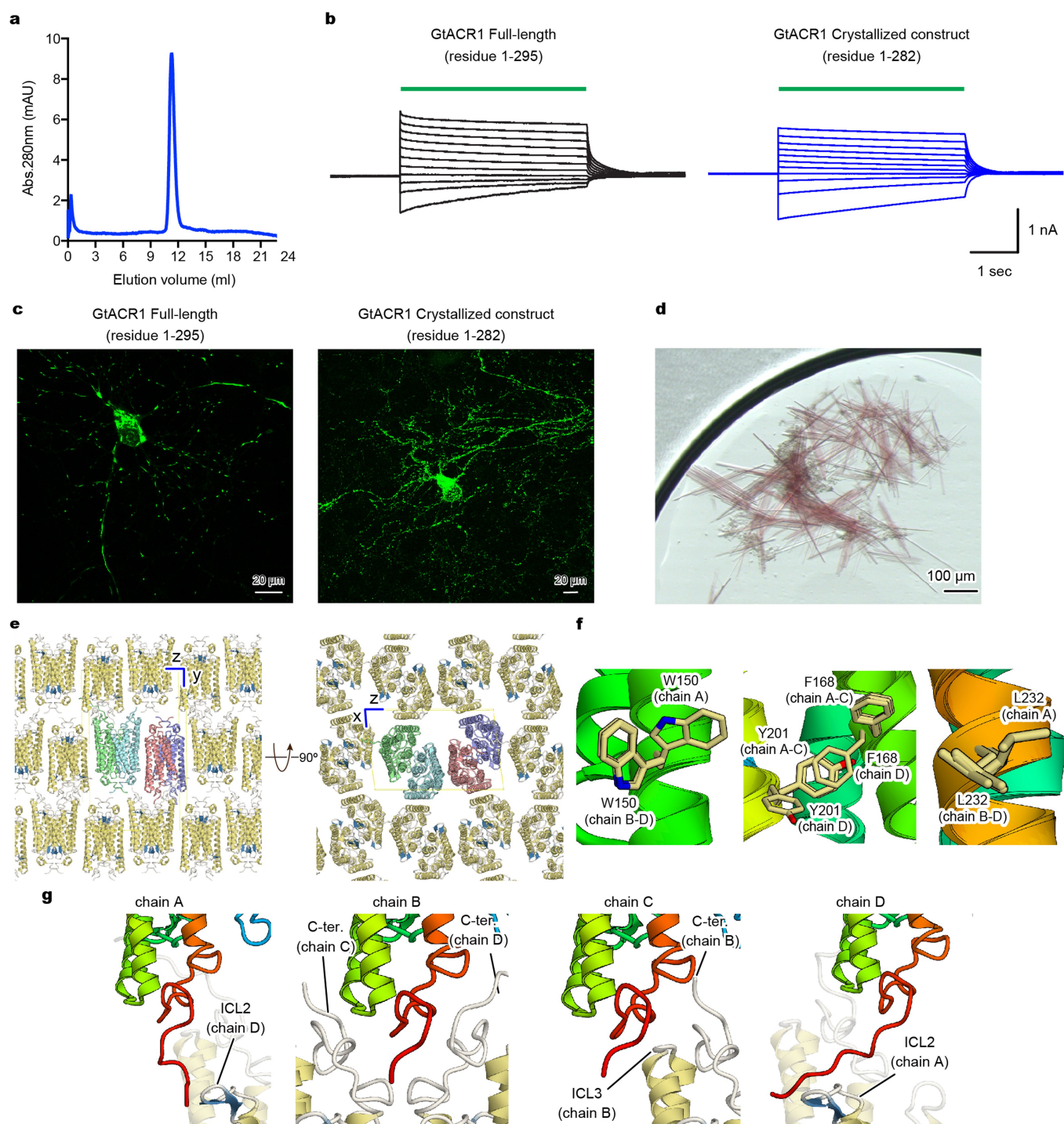
Light-induced difference FTIR spectroscopy. Wild-type and D234N mutant *GtACR1* were reconstituted into a mixture of POPE and POPG (molar ratio = 3:1) with a protein-to-lipid molar ratio of 1:30 by removing DDM with Bio-Beads (SM-2, Bio-Rad). The reconstituted samples were washed three times with buffers at pH 5.0 (2 mM citrate-NaOH), pH 7.0 (2 mM HEPES-NaOH) or pH 9.0 (2 mM borate-NaOH) with 1 mM NaCl. The pellet was re-suspended in the same buffer, with the concentration adjusted to 1.7 mg ml⁻¹. A 60 µl aliquot was placed onto a BaF₂ window and air-dried. FTIR spectroscopy was applied to the films hydrated with 1 µl H₂O at 77 K, 170 K and 200 K as described previously⁵⁸. In brief, the sample was placed in an Oxford DN-1704 cryostat mounted in the Bio-Rad FTS-40 spectrometer (instrumental resolution of FTIR is 2 cm⁻¹). For the formation of photo-intermediates at 77 K, samples were illuminated at 500 nm (interference filter) from a 1-kW halogen-tungsten lamp for 2 min and photo-reversed with >600 nm light (R-62 cut-off filter, Toshiba) for 1 min. For formation of photo-intermediates at 170 K and 200 K, samples were illuminated with >500 nm light (Y-52 cut-off filter, Toshiba) for 1 min. For each measurement of FTIR spectroscopy, 256 interferograms were accumulated; 40 identical recordings at 77 K and 7 identical recordings at 170 K and 200 K were averaged.

Measurement of UV absorption spectra. Protein absorbance spectra were measured with an Infinite M1000 microplate reader (Tecan Systems Inc.) using 96 well plates (ThermoFisher scientific). The *GtACR1* samples were suspended in a buffer containing 100 mM NaCl, 0.05% DDM, 0.01% CHS, and 20 mM sodium citrate, sodium acetate, sodium cacodylate, HEPES, Tris, CAPSO or CAPS. pH was adjusted from 4 to 10 by the addition of NaOH or HCl. Recorded spectra value was averaged from 20 measurements from a single session.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

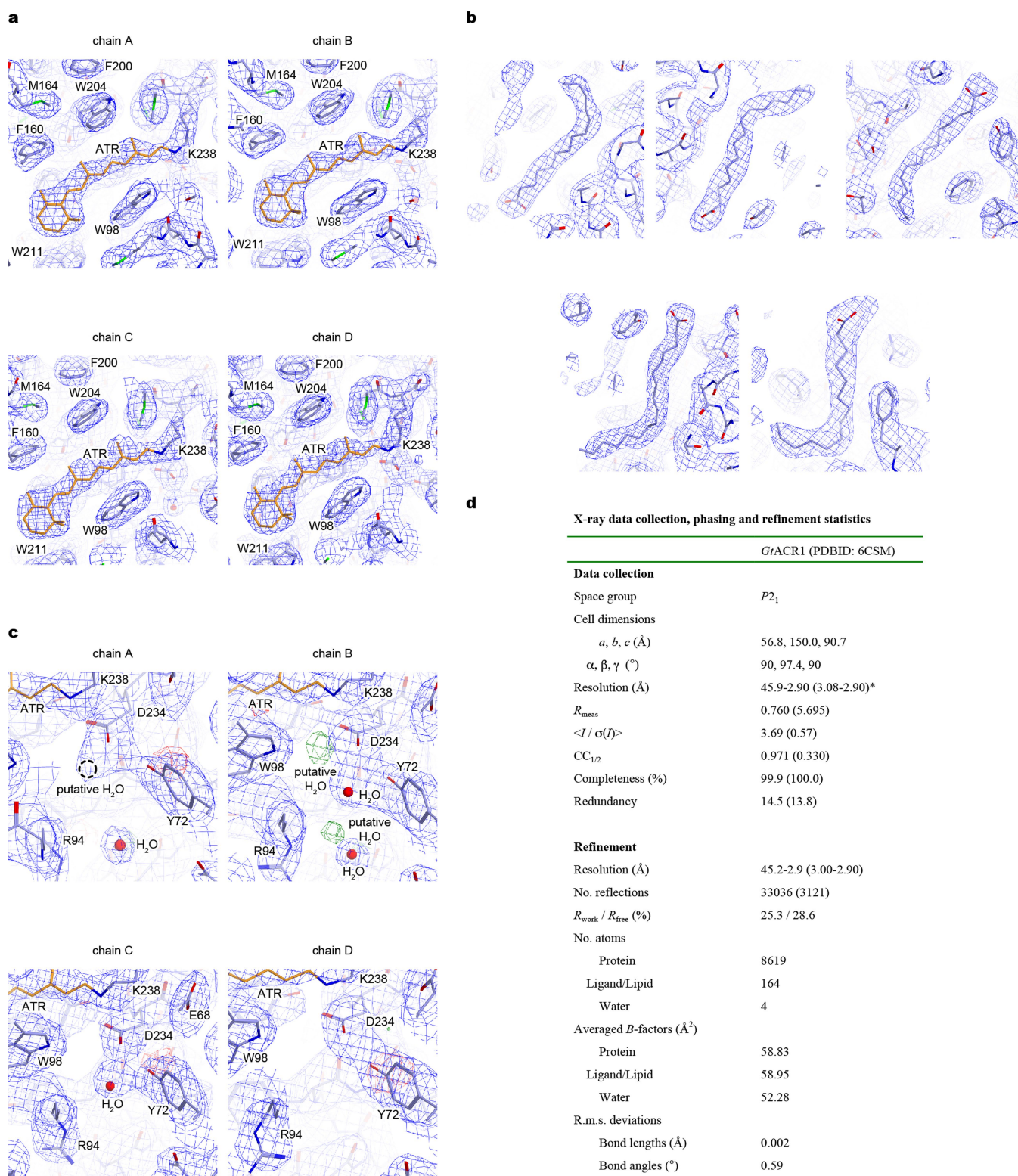
Data availability. The protein coordinate and atomic structure factor have been deposited in the Protein Data Bank (PDB) under accession number 6CSM. The raw diffraction images have been deposited in the SBGrid Data Bank repository (ID: 569). All other data are available from the corresponding authors upon reasonable request.

52. Yamashita, K., Hirata, K. & Yamamoto, M. KAMO: towards automated data processing for microcrystals. *Acta Crystallogr. D* **74**, 441–449 (2018).
53. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
54. Murshudov, G. N. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D* **67**, 355–367 (2011).
55. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
56. DiMaio, F. Advances in Rosetta structure prediction for difficult molecular-replacement problems. *Acta Crystallogr. D* **69**, 2202–2208 (2013).
57. Tanimoto, T., Furutani, Y. & Kandori, H. Structural changes of water in the Schiff base region of bacteriorhodopsin: proposal of a hydration switch model. *Biochemistry* **42**, 2300–2306 (2008).
58. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
59. Luecke, H., Schobert, B., Richter, H. T., Cartailier, J. P. & Lanyi, J. K. Structure of bacteriorhodopsin at 1.55 Å resolution. *J. Mol. Biol.* **291**, 899–911 (1999).
60. Kato, H. E. et al. Structural basis for Na⁺ transport mechanism by a light-driven Na⁺ pump. *Nature* **521**, 48–53 (2015).
61. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).
62. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucl. Acids Res.* **42**, W320–W324 (2014).



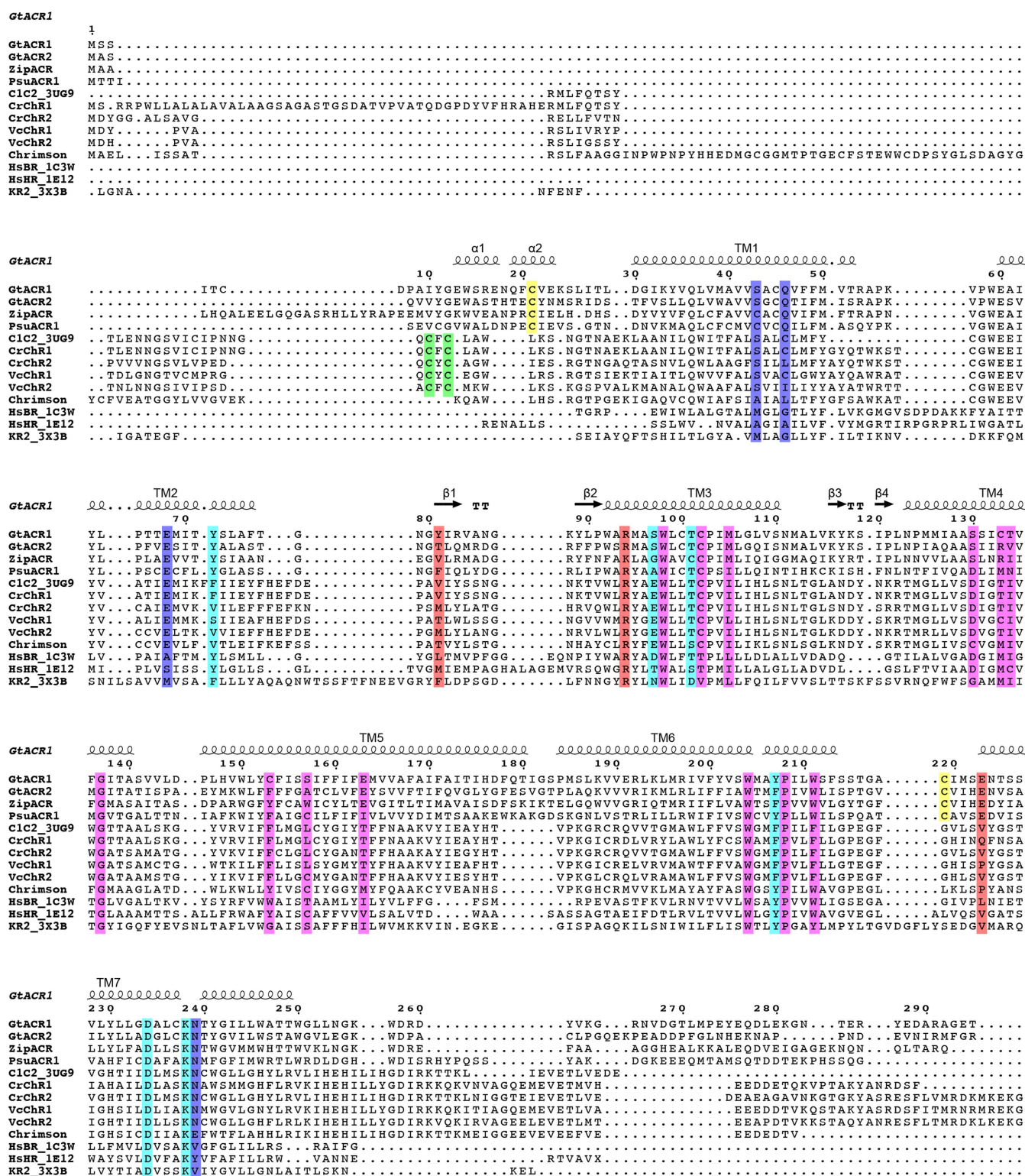
Extended Data Fig. 1 | Crystallography. **a**, Size exclusion chromatogram of the purified *GtACR1* protein used for crystallography. Similar results were seen in more than 20 independent experiments. **b**, Electrophysiology of full-length *GtACR1* (left) and the final crystallization construct (right); whole-cell voltage-clamp recordings in five cells held at -70 mV, with 513 nm light at 1.0 mW mm $^{-2}$ irradiance delivered with timing as shown with green-coloured bars, while cells were held at resting potentials from -95 mV (lowest trace) to $+5$ mV (uppermost trace) in steps of 10 mV. Similar results were seen in 3–5 cells from each group, and no significant difference was seen in resting potential, input resistance, reversal potential

or photocurrent magnitude. **c**, Confocal images of cultured hippocampal neurons expressing full-length *GtACR1* (left) and the final crystallization construct (right). Similar results were seen in more than five cells from 3–5 coverslips. Note the markedly reduced aggregation of the truncated construct. **d**, Crystals of *GtACR1*. Similar crystals were generated in more than 200 experiments. **e**, Lattice packing of *GtACR1* crystals, viewed parallel to the x axis (left) and the y axis (right). **f**, Different amino acid configurations at different chains within the asymmetric unit of *GtACR1*. **g**, C-terminal interactions among different chains within the asymmetric unit of *GtACR1*.



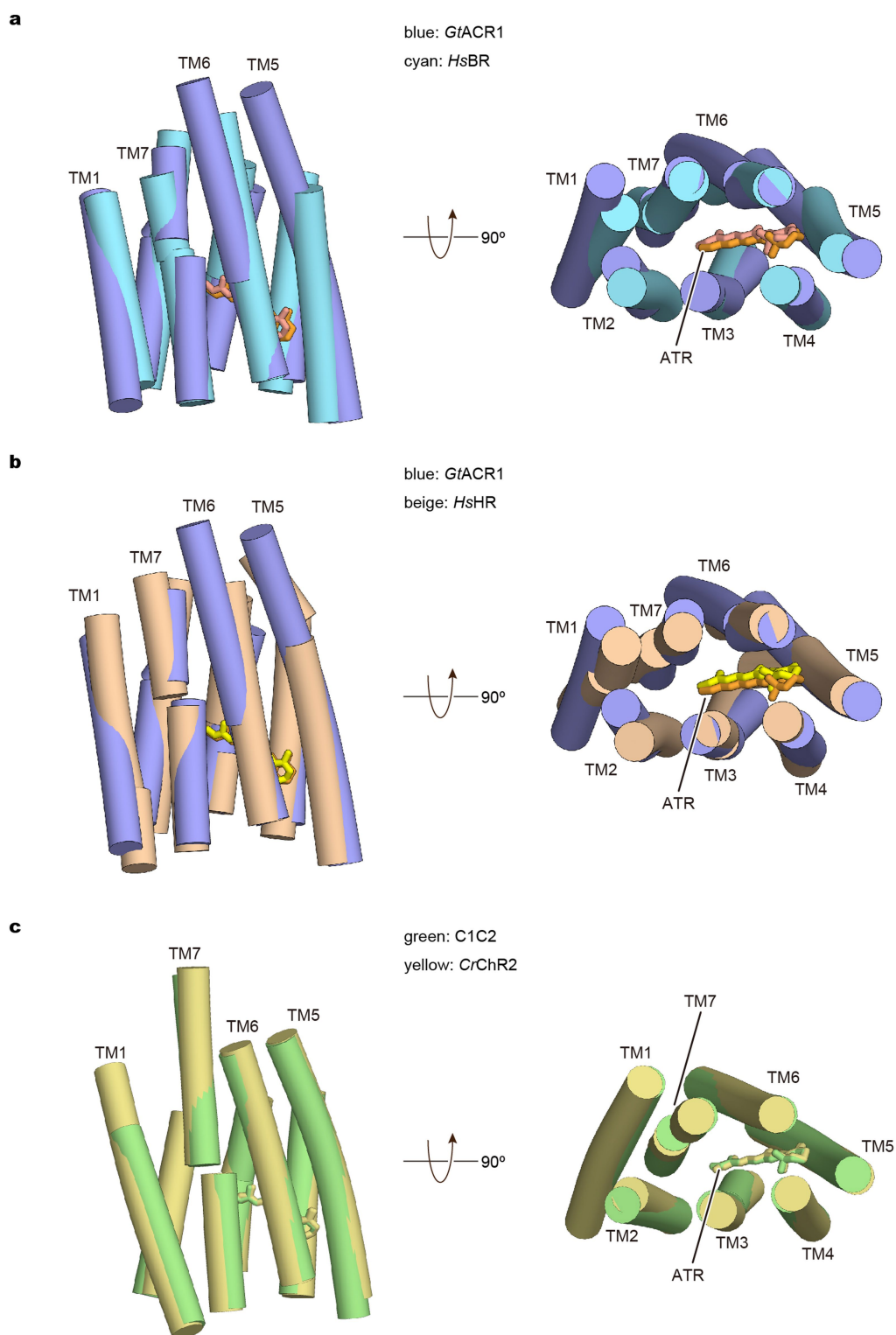
Extended Data Fig. 2 | Structural analysis of *GtACR1*. **a**, $2F_o - F_c$ maps (blue mesh, contoured at 1σ) for the retinal-binding pockets of chains A–D. **b**, $2F_o - F_c$ maps (blue mesh, contoured at 1σ) for the lipid molecules. **c**, $2F_o - F_c$ maps (blue mesh, contoured at 1σ) and $F_o - F_c$ maps (green and red meshes, contoured at 3.0σ and -3.0σ , respectively)

for the Schiff base region of chains A–D. Water molecules are shown as red spheres. **d**, Table describing data collection and refinement statistics of *GtACR1*. Dataset was collected from 80 crystals. Values in parentheses are for the highest-resolution shell.



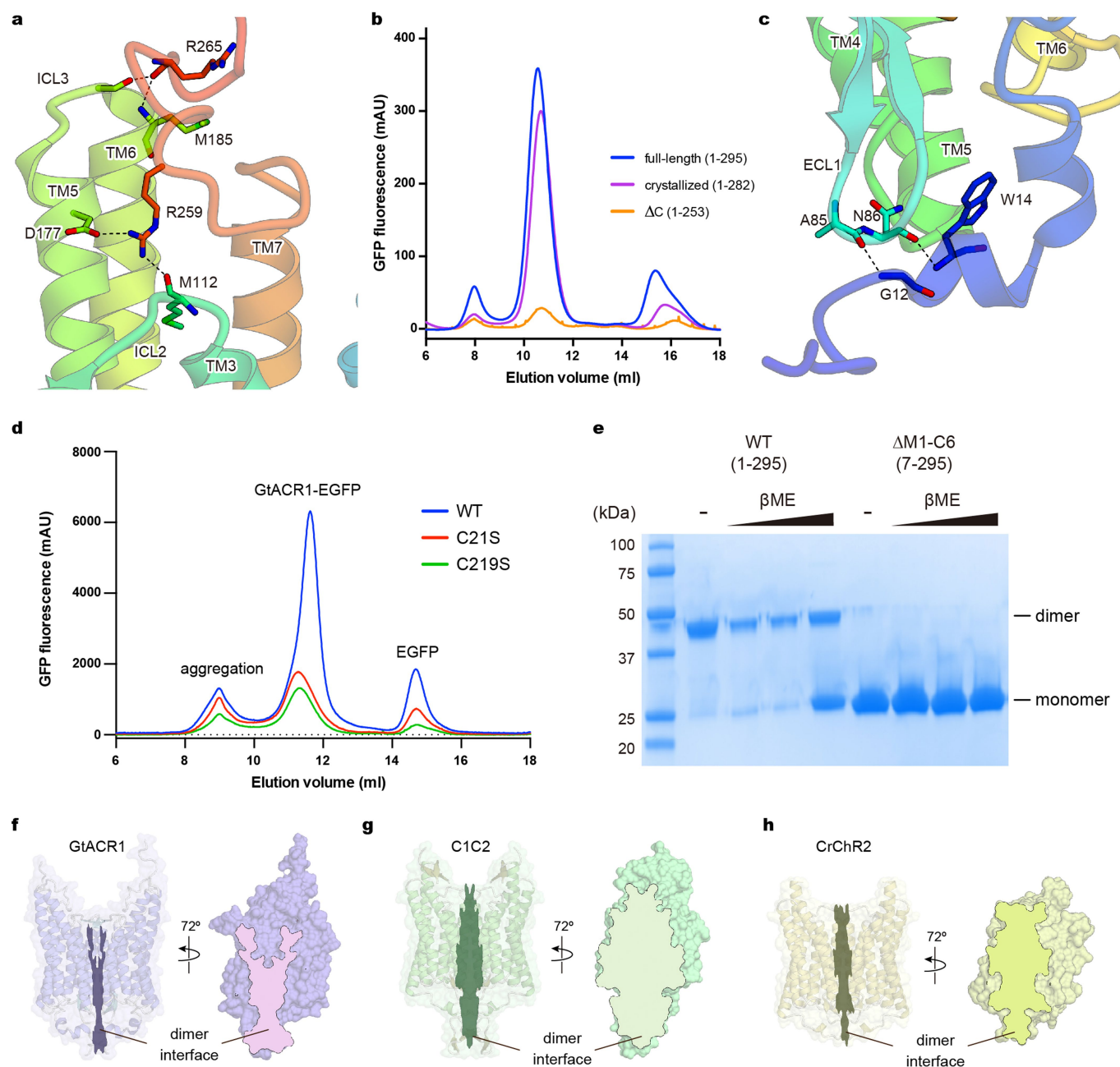
Extended Data Fig. 3 | Structure-based sequence alignment of microbial opsin genes. The sequences are GtACR1 (GenBank accession AKN63094.1), GtACR2 (GenBank AKN63095.1), ZipACR (GenBank APZ76709.1), PsuACR1 (GenBank ID: KF992074.1), the chimaeric channelrhodopsin between CrChR1 and CrChR2 (C1C2, PDB code 3UG9)³⁴, CrChR1 (GenBank 15811379), CrChR2 (GenBank 158280944), ChR1 from *Volvox carteri* (VcChR1, UniProtKB B4Y103), ChR1 from *V. carteri* (VcChR2, UniProtKB ID: B4Y105), Chrimson (GenBank ID: AHH02126.1), ChR from *Tetraselmis striata* (TsChR, GenBank ID:

KF992089.1), HsBR (PDB code 1C3W)⁵⁹, HsHR (PDB code 1E12)⁴⁸, and *Krokinobacter eikastus* rhodopsin 2 (KR2, PDB code 3X3B)⁶⁰. The sequence alignment was created using PROMALS3D⁶¹ and ESPrnt³ servers. Secondary structure elements for GtACR1 are shown as coils and arrows. ‘TT’ represents turns. Cysteine residues forming intermolecular and intramolecular disulfide bridges are highlighted in green and yellow, respectively. The residues of retinal-binding pockets are coloured pink. The residues in the Schiff base region are coloured cyan. The residues forming the ECS2 and CCS are coloured orange and blue, respectively.



Extended Data Fig. 4 | Structural comparison among *Gt*ACR1, *Hs*BR, *Hs*HR, C1C2 and *Cr*ChR2. a, b, Side view and extracellular view of the superimposed transmembrane regions of *Gt*ACR1 (blue) and *Hs*BR (cyan) (a), *Gt*ACR1 (blue) and *Hs*HR (beige) (b), C1C2 (green) and

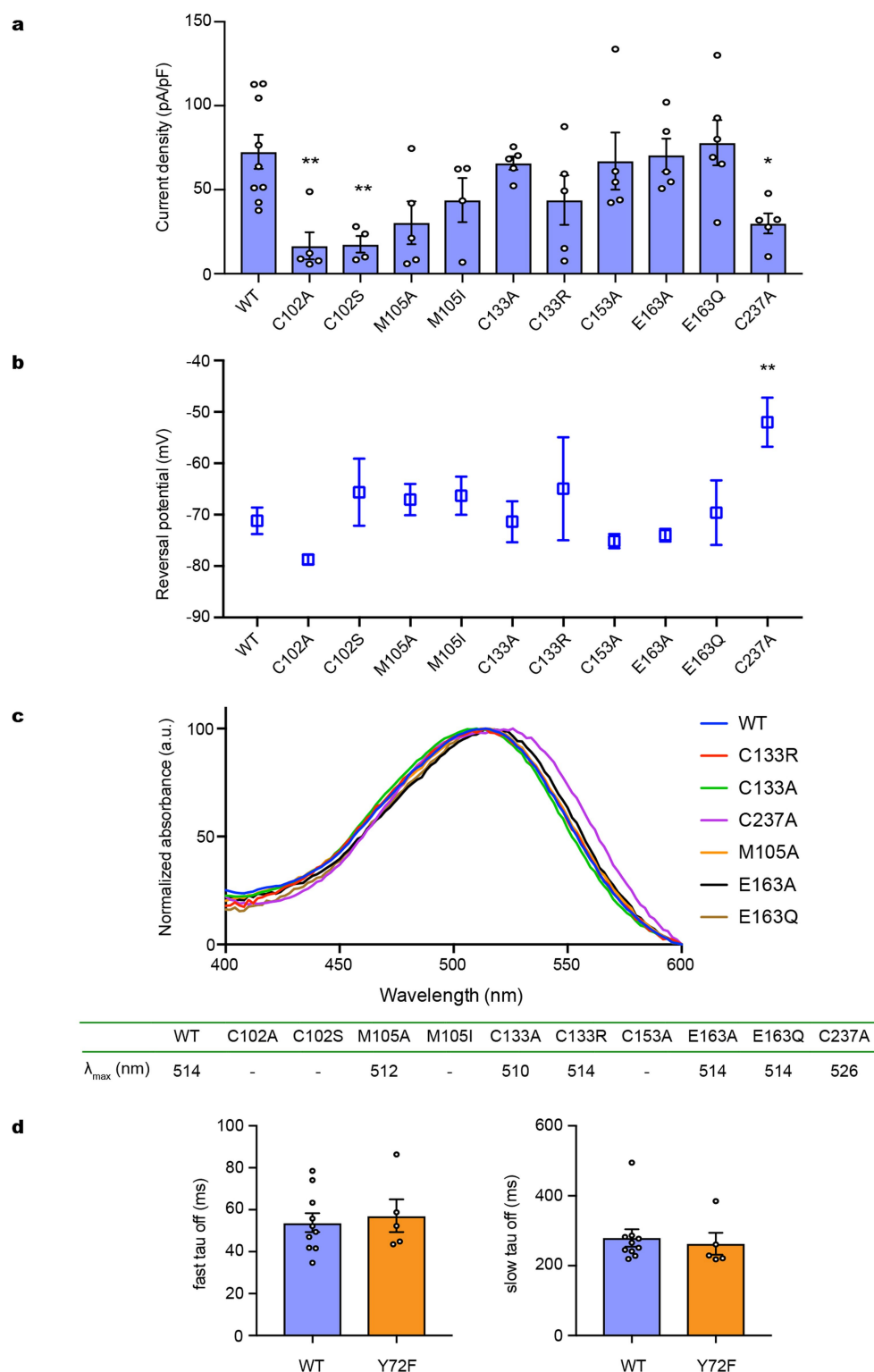
*Cr*ChR2 (yellow) (c). The ATRs are shown as stick models, and are coloured orange (*Gt*ACR1), salmon (*Hs*BR), light-yellow (*Hs*HR), green (C1C2) and yellow (*Cr*ChR2).



Extended Data Fig. 5 | Interactions between N- and C-terminal regions and the 7-TM domain.

a, Interactions between the C-terminal region and the 7-TM domain. Hydrogen bonds are shown by dashed lines. **b**, Fluorescent size-exclusion chromatography traces of the full-length *GtACR1* (1–295), the crystallized construct (1–282), and the C-terminal truncated construct (Δ C: 1–253), showing possible importance of the C terminus in proper folding and/or stability. Similar results were observed in three independent experiments. **c**, Interactions between the N-terminal region and the ECL1. Hydrogen bonds are shown by dashed lines. **d**, Fluorescent size-exclusion chromatography traces of wild-type and C-to-S mutants of *GtACR1*. Labels indicate estimated elution positions of the aggregate, *GtACR1*-eGFP, and free eGFP; C-to-S mutants show

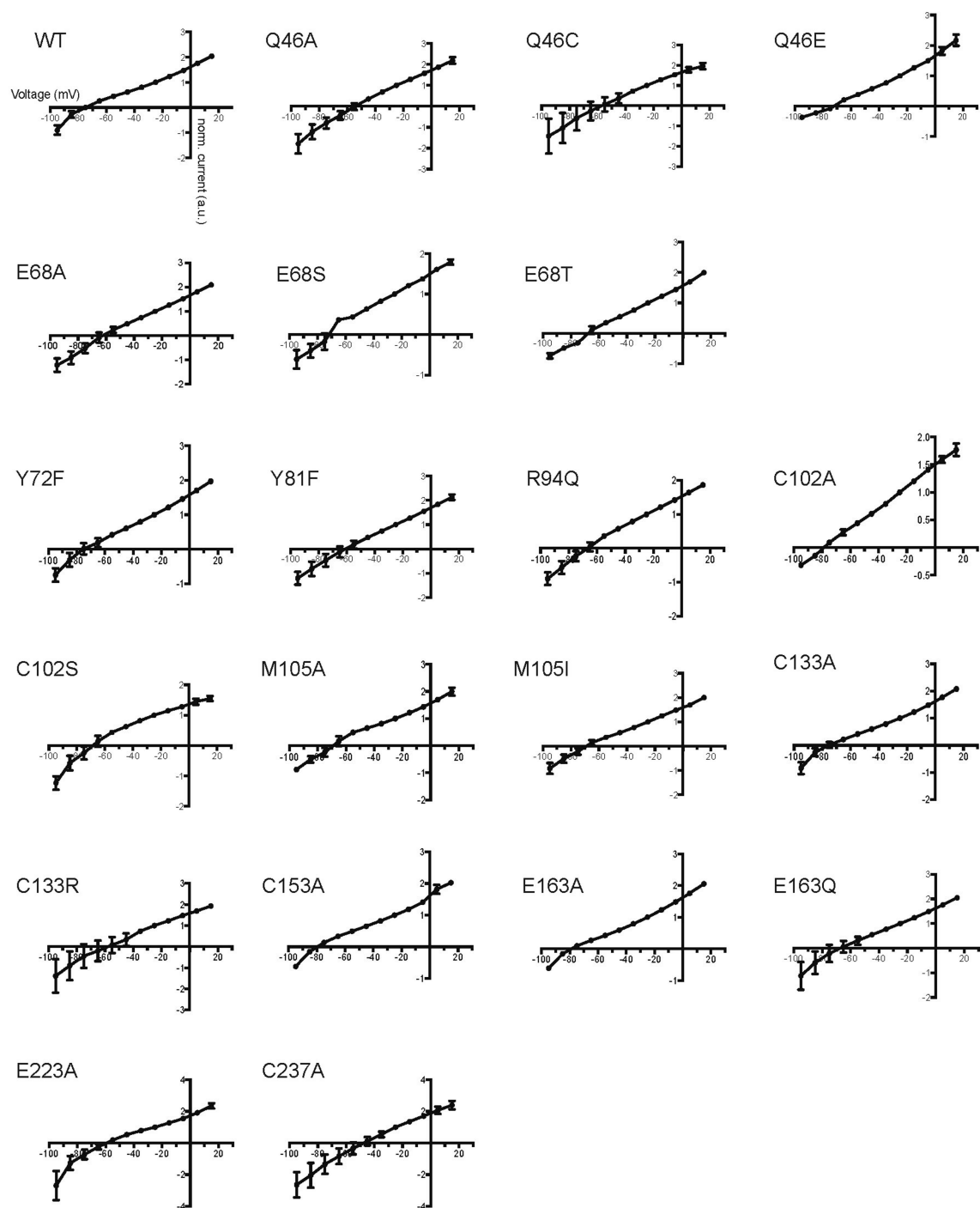
decreased ($<1/3$) expression compared to the wild type. Similar results were observed in three independent experiments. **e**, Stained SDS-PAGE gel image of wild-type and N-terminal 6-amino-acid-truncated *GtACR1* in the presence and absence of reducing reagent (β -mercaptoethanol); the wild type runs as a mixer of monomer and dimer in β -mercaptoethanol, whereas N-terminal-truncated *GtACR1* stays monomeric even in the absence of β -mercaptoethanol. This experiment was performed once, but similar experiments with different concentrations of β -mercaptoethanol were performed three times, all with similar results. **f–h**, Dimer interfaces of *GtACR1* (**f**), C1C2 (**g**) and CrChR2 (**h**) viewed at two angles from the side; note reduced interface area (outlined) for *GtACR1*. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 6 | Conductances, reversal potentials, absorption spectra and kinetics of wild-type *GtACR1* and mutants.

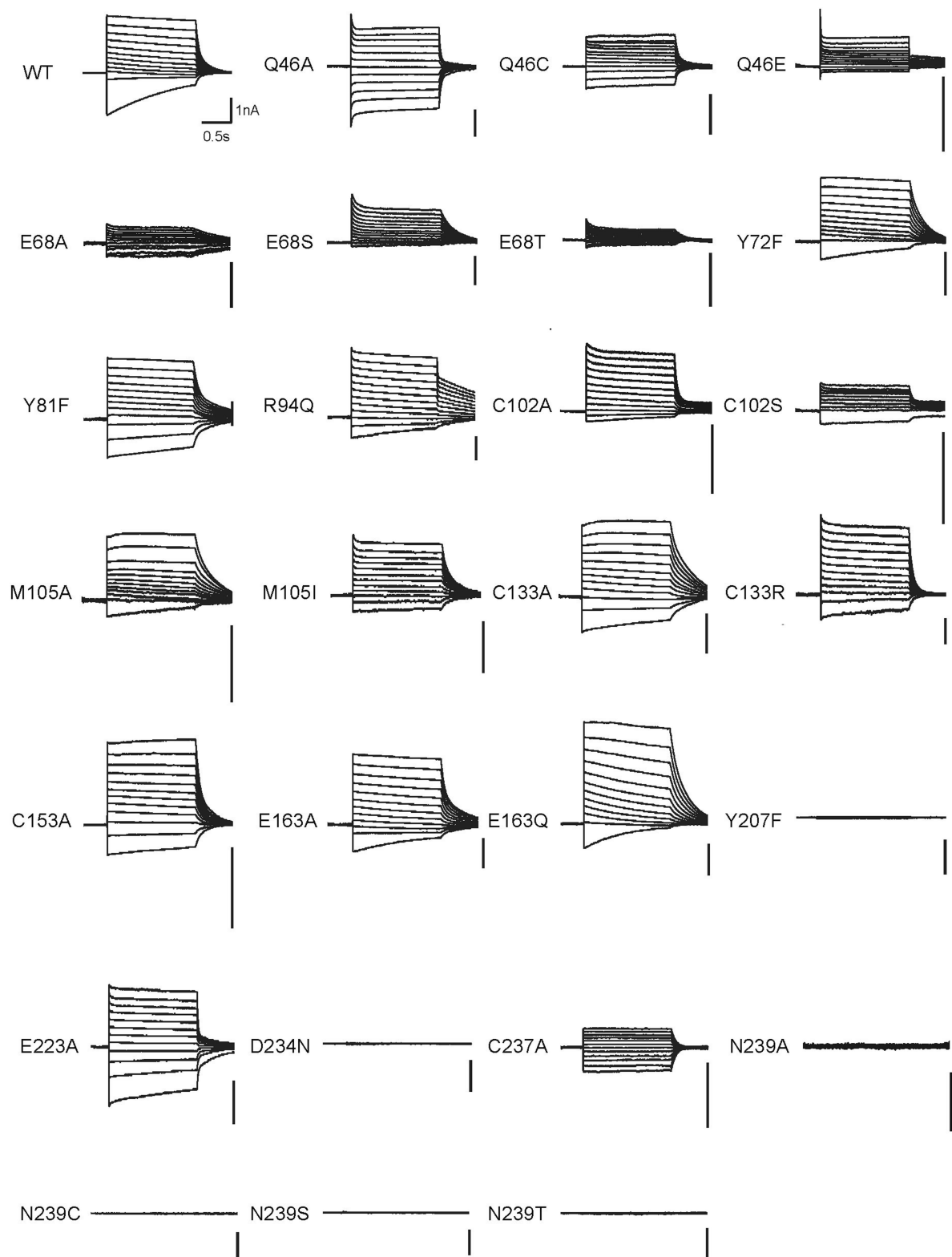
a–c, Photocurrents (**a**), reversal potentials (**b**) and absorption spectra (**c**) of wild-type *GtACR1* and ten mutants of the retinal-binding pocket. λ_{\max} values are listed in the table (**c**, bottom). Photocurrents are measured in whole-cell voltage-clamp recordings held at -70 mV, with 513 nm light at 1.0 mW mm $^{-2}$ irradiance. Data are mean and s.e.m.; $n = 9$ for WT, 6 for E163Q, 5 for C102A, M105A, C133A, C133R, C153A, E163A and C237A, and 4 for the rest. $*P < 0.05$, $**P < 0.01$, one-way ANOVA followed by Dunnett's test. Reversal potentials are measured with identical light

stimulation while cells were held at resting potentials from -95 mV to $+15$ mV in steps of 10 mV. Data are mean and s.e.m. $n = 10$ for WT and C237A, 6 for E163A and E163Q, 5 for C102A, M105A, C133A and C153A, and 4 for the rest. $**P = 0.0022$, one-way ANOVA followed by Dunnett's test. Spectra measurement was performed in two independent trials, with wild type as a positive control. **d**, Comparison of fast closing (left) and slow closing (right) coefficients of wild-type and Y72F mutant *GtACR1*. Data are mean and s.e.m. $n = 10$ for WT and 5 for Y72F. $P = 0.7$ for both graphs, two-tailed t -test.

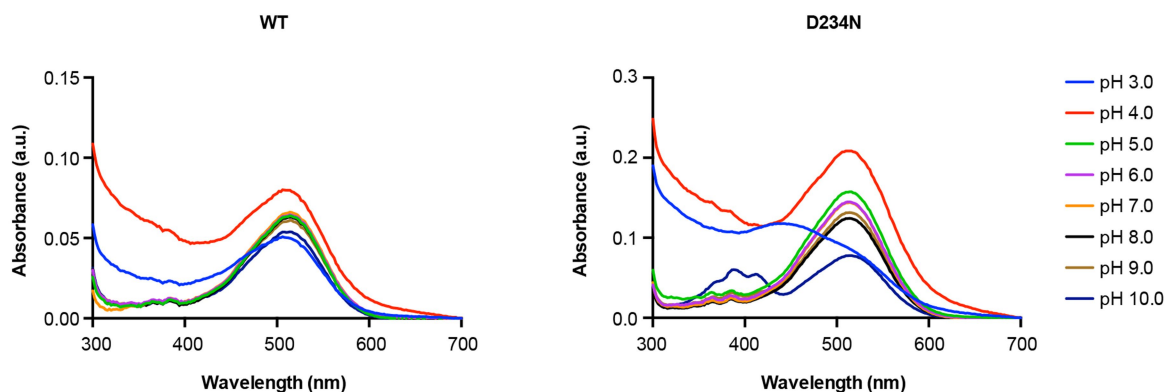


Extended Data Fig. 7 | Current-voltage (I - V) relationships of wild-type *GtACR1* and mutants. The I - V relationship between -95 mV and $+15$ mV was determined from the single current amplitude at the indicated potentials. Each measurement is normalized to the current amplitude

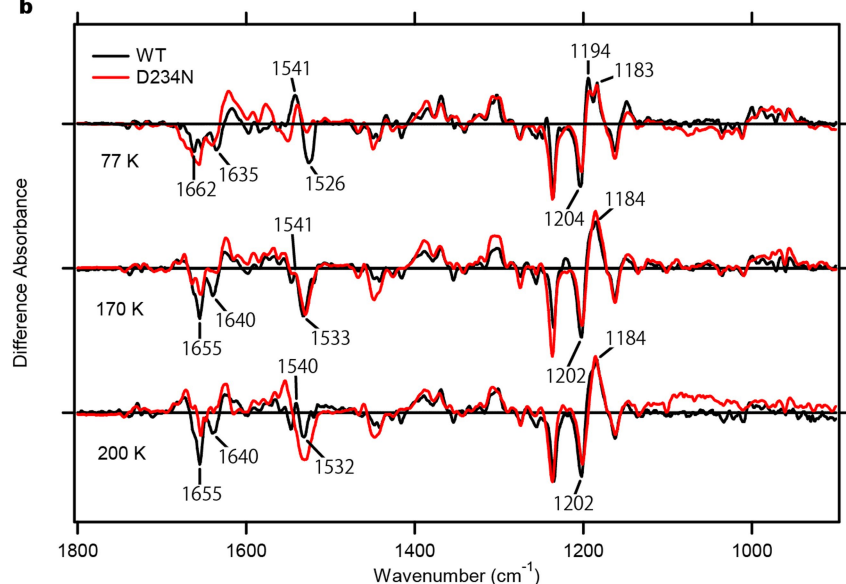
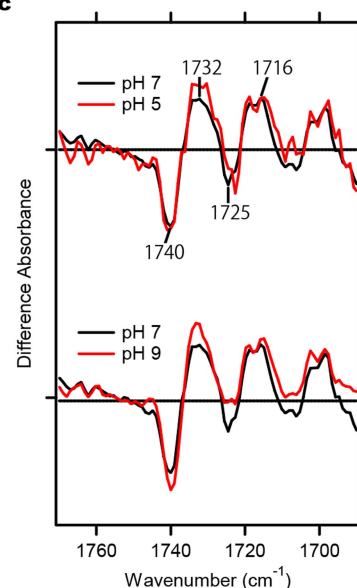
measured at -25 mV. Data are mean and s.e.m. $n = 10$ for WT and C237A, 8 for E223A, 6 for Q46C, E163A and E163Q, 4 for E68S, E68T, C102S and M105I, and 5 for the rest.



Extended Data Fig. 8 | Representative traces of the I - V measurement of wild-type *GtACR1* and mutants. Voltage clamp traces corresponding to the I - V relationships in Extended Data Fig. 7 between -95 mV and $+15$ mV.

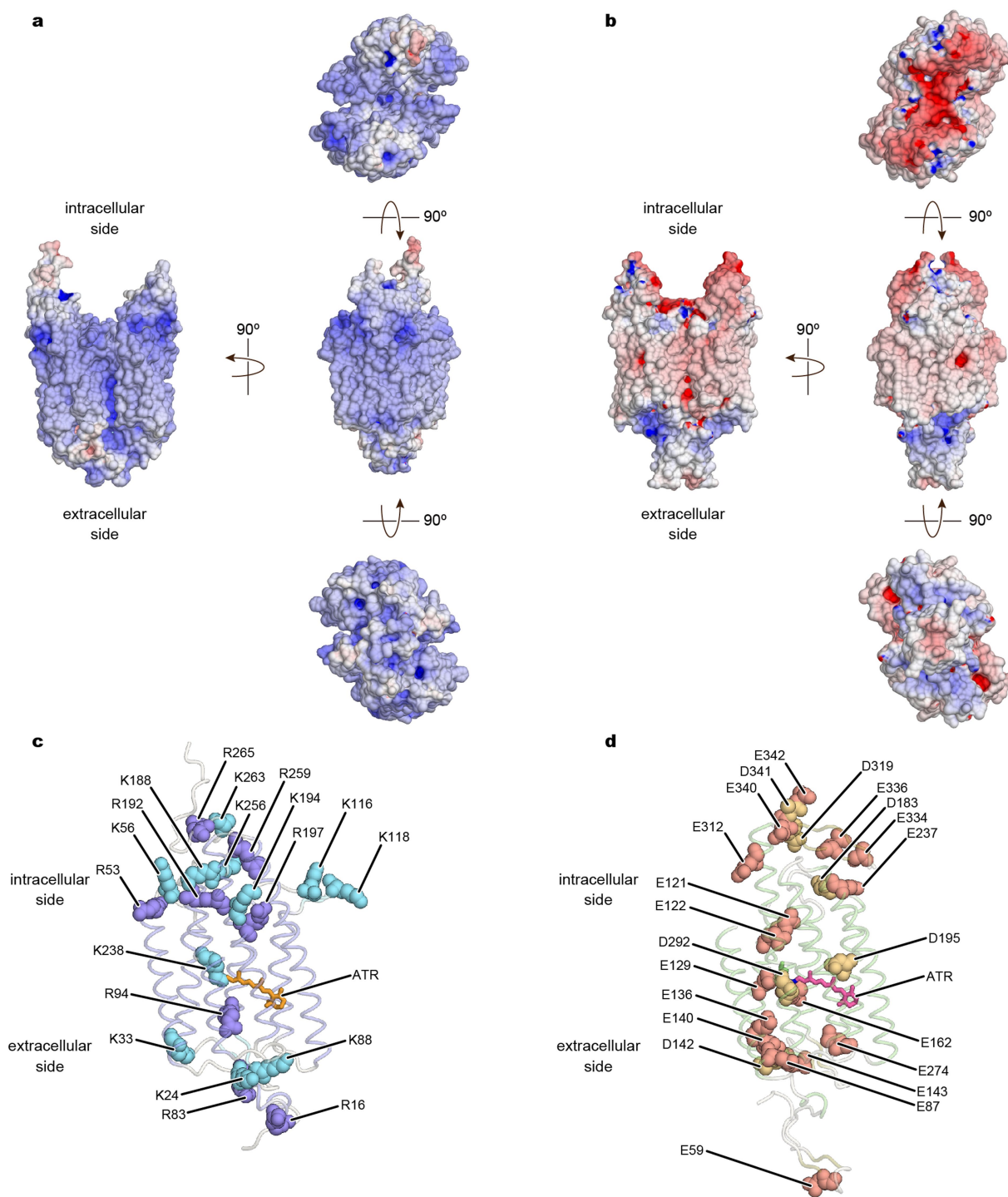
a

	pH3	pH4	pH5	pH6	pH7	pH8	pH9	pH10
λ_{\max} (nm)	WT	506	506	514	514	514	514	514
	D234N	442	514	514	514	514	514	514

b**c**

Extended Data Fig. 9 | Spectroscopic characterization of wild-type *GtACR1* and the D234N mutant. **a**, Absorption spectra of wild-type *GtACR1* (top left) and the D234N mutant (top right) measured from pH 3.0 to 10.0. The λ_{\max} value at each pH is listed in the table (bottom). **b**, Difference FTIR spectra of wild-type *GtACR1* and the D234N mutant

measured at 77 K, 170 K and 200 K. **c**, Difference FTIR spectra of wild-type *GtACR1* in the 1,690–1,770 cm^{-1} region measured at pH 5.0, 7.0 and 9.0. Forty identical recordings at 77 K and seven identical recordings at 170 K and 200 K were averaged.



Extended Data Fig. 10 | Comparison of surface electrostatic potential of *GtACR1* and *C1C2*. **a, b**, Electrostatic potential surfaces of *GtACR1* (**a**) and *C1C2* (**b**) viewed from four angles. The surface is coloured on the basis of the electrostatic potential contoured from -15 kT (red) to $+15$ kT

(blue). **c, d**, Representation of positively charged amino acids (lysine and arginine residues) in *GtACR1* (**c**), and negatively charged amino acids (aspartate and glutamate residues) in *C1C2* (**d**).

Structural mechanisms of selectivity and gating in anion channelrhodopsins

Hideaki E. Kato^{1,2,13*}, Yoon Seok Kim^{3,4,5,13}, Joseph M. Paggi^{6,7}, Kathryn E. Evans^{3,4,5}, William E. Allen^{3,4,5}, Claire Richardson⁶, Keiichi Inoue^{2,10,11}, Shota Ito¹⁰, Charu Ramakrishnan^{3,4,5}, Lief E. Fenno^{3,4,5}, Keitaro Yamashita¹², Daniel Hilger¹, Soo Yeun Lee^{3,4,5}, Andre Berndt^{3,4,5}, Kang Shen^{8,9}, Hideki Kandori^{10,11}, Ron O. Dror^{6,7}, Brian K. Kobilka¹ & Karl Deisseroth^{3,4,5*}

Both designed and natural anion-conducting channelrhodopsins (dACRs and nACRs, respectively) have been widely applied in optogenetics (enabling selective inhibition of target-cell activity during animal behaviour studies), but each class exhibits performance limitations, underscoring trade-offs in channel structure–function relationships. Therefore, molecular and structural insights into dACRs and nACRs will be critical not only for understanding the fundamental mechanisms of these light-gated anion channels, but also to create next-generation optogenetic tools. Here we report crystal structures of the dACR iC⁺⁺, along with spectroscopic, electrophysiological and computational analyses that provide unexpected insights into pH dependence, substrate recognition, channel gating and ion selectivity of both dACRs and nACRs. These results enabled us to create an anion-conducting channelrhodopsin integrating the key features of large photocurrent and fast kinetics alongside exclusive anion selectivity.

Anion selectivity within the broad family of light-gated ion channels¹ was initially created by transformation, guided by crystal structures², of natural cation-conducting channelrhodopsins (CCRs) into dACRs^{3–6}. Anion selectivity was subsequently found to have evolved naturally in the nACRs of certain cryptophyte algae⁷. Both classes of anion-conducting channelrhodopsin (ACR) have proved to be useful in optogenetics—that is, for enabling reversible silencing of specific neurons with sufficient potency to modulate animal behaviour, beginning⁶ with the dACR iC⁺⁺.

Befitting their provenance from CCRs that achieved versatile applicability in optogenetics⁸, in part through engineered gating spanning around 6 orders of magnitude from millisecond-scale control of fast-spiking cells to step-like bistable modulation^{1,3,6,9,10}, dACRs offer a wider range of kinetics relevant to neuroscience than nACRs. On the other hand, nACRs exhibit larger photocurrents (despite high anion selectivity)⁷. Bringing important features such as speed and large photocurrents together in a single opsin has been hampered by a lack of structural understanding.

The structure of an nACR, from *Guillardia theta* (GtACR1; reported in the accompanying paper¹¹, provided high-resolution perspective on anion selectivity. However, complementary structural information on dACRs would greatly accelerate development of next-generation tools for neuroscience¹². Moreover, formal confirmation of the basis for anion selectivity and channel gating would require detailed structural information and redesign of dACRs. We therefore solved multiple structures of the dACR iC⁺⁺^{6,13–16} and used resulting insights to create a new class of ACR functionality.

Structural and functional characterization of iC⁺⁺

iC⁺⁺ was made from a CCR (C1C2) via 10 mutations^{3,6} in transmembrane helices (TM) 1, 2, 3 and 7 (Extended Data Fig. 1a). To functionally characterize iC⁺⁺ states, we performed photocycle analysis via flash

photolysis, observing three spectroscopically distinguishable intermediates (K, M and O). The K intermediate had the longest lifetime of any channelrhodopsin ($\tau = 2.6 \pm 0.1$ ms; Extended Data Fig. 1b–d), and channel closing appeared to be coupled to decay of the M intermediate (Extended Data Fig. 1e). The photocycle of iC⁺⁺ can therefore be distinguished from those of nACRs¹⁷; for example, GtACR1 has 5 intermediate states (K, L, M, N, O) with a much shorter K lifetime (microsecond scale) than iC⁺⁺ (Supplementary Discussion). Gating of nACRs may be regulated by two kinetic processes (coupled slow-opening–fast-closing and fast-opening–slow-closing), with fast and slow closing coupled to decay of L and M intermediates, respectively¹⁷; key differences between iC⁺⁺ and GtACR1 photocycle architecture therefore provide opportunities to explore underlying structural mechanisms.

We obtained crystal structures of iC⁺⁺ at pH 6.5 and 8.5 (at 3.2 and 2.9 Å resolution, respectively; Fig. 1a, Extended Data Fig. 1f). To facilitate crystallogensis, an N-linked glycosylation site was mutated (N61Q), and crystallization was achieved within the lipidic cubic phase composed of monolein (Extended Data Fig. 2a). The two structures were almost identical (root mean square deviation (r.m.s.d.) of 0.31 Å over all C_α atoms; Extended Data Fig. 2b). Like other channelrhodopsins¹, iC⁺⁺ forms a dimer; each monomer is composed of an N-terminal extracellular domain (residues 51–84), 7 transmembrane domains (residues 85–316), and a C-terminal β -sheet domain (residues 317–342) (Fig. 1a). The 10 mutated residues were well resolved (Fig. 1a, Extended Data Fig. 3a), and we could model two putative water molecules into the electron-density map that were stable in molecular dynamics simulations (Extended Data Fig. 3b).

Previous studies had found differences in pH dependence among certain nACRs and dACRs^{3,6,7}; in nACRs such as GtACR1, channel activity is insensitive to extracellular pH, but in dACRs photocurrent can decrease under alkaline conditions (Fig. 1b). Comparing the

¹Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA. ²PRESTO, Japan Science and Technology Agency, Kawaguchi, Japan. ³Department of Bioengineering, Stanford University, Stanford, CA, USA. ⁴Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA. ⁵Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. ⁶Department of Computer Science, Stanford University, Stanford, CA, USA. ⁷Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA. ⁸Department of Biology, Stanford University, Stanford, CA, USA. ⁹Howard Hughes Medical Institute, Stanford University, Stanford, CA, USA. ¹⁰Department of Life Science and Applied Chemistry, Nagoya Institute of Technology, Nagoya, Japan. ¹¹OptoBioTechnology Research Center, Nagoya Institute of Technology, Nagoya, Japan. ¹²RIKEN SPring-8 Center, Hyogo, Japan. ¹³These authors contributed equally: Hideaki E. Kato; Yoon Seok Kim. *e-mail: hekato@stanford.edu; deissero@stanford.edu

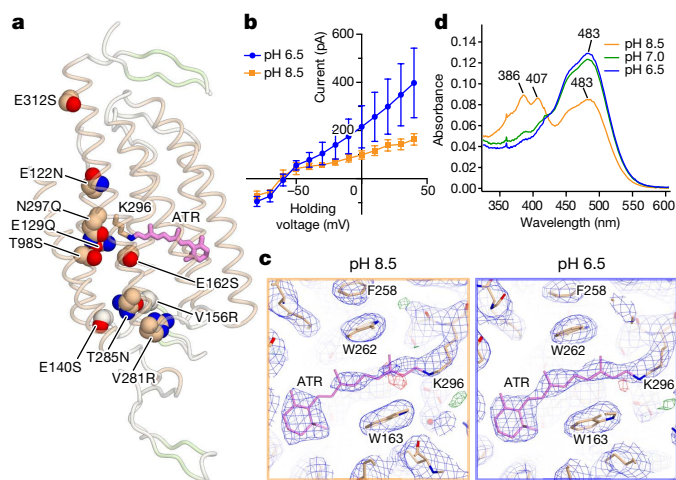


Fig. 1 | Structures of iC++ and insights into pH dependence. **a**, Crystal structure of iC++ at pH 8.5. iC++ mutations are shown with sphere models. **b**, pH-dependent photocurrents of iC++ at pH_{ext} 6.5 (blue) and 8.5 (orange). Data are mean \pm s.e.m. of 5 cells. **c**, Retinal-binding pocket (RBP) of iC++ at pH 8.5 and 6.5; $2F_o - F_c$ maps (blue mesh, contoured at 1σ) and $F_o - F_c$ maps (green and red meshes, contoured at 3σ and -3σ , respectively) are shown. **d**, iC++ absorbance spectra at pH 6.5, 7.0 and 8.5.

structures at pH 6.5 and 8.5 revealed that occupancy of the all-*trans*-retinal (ATR) chromophore site markedly differed. Although resolution and quality of the overall electron-density map at pH 8.5 was slightly better than at pH 6.5, the local electron density of ATR was lower (Fig. 1c), indicating a structural difference. In calculating mean electron densities of the overall protein and of ATR using the diffraction to 3.2 \AA , mean electron density of the overall structure was indeed higher, but that of ATR was lower, at pH 8.5 (Extended Data Fig. 2c), suggesting that occupancy of ATR was lower at pH 8.5, consistent with decreased colouration of the crystals (Extended Data Fig. 2a). We also measured absorption spectra of iC++ at different pH values and observed additional peaks at 386 nm and 407 nm for pH values above 7.5, suggesting that the Schiff base is deprotonated and ATR is hydrolyzed under alkaline conditions (Fig. 1d, Extended Data Fig. 2d, e). Reduction of electronegative potential by introduced mutations such as E129Q and E162S could destabilize the Schiff base and allow hydrolysis of ATR at alkaline pH, thus decreasing dACR activity.

Extracellular constriction sites of ACRs

We next analysed the ion-conduction pathway, which has implications for both basic biology and optogenetics. All structurally resolved channelrhodopsins contain two extracellular vestibules (EV1 and EV2) (Extended Data Fig. 4). In C1C2, EV1 is occluded at the extracellular constriction site-1 (ECS1) by hydrogen-bonding interactions between Gln95 and Glu140, whereas EV2 extends to the central constriction site (CCS)^{2,11}. In *GtACR1*, EV2 is occluded at extracellular constriction site-2 (ECS2) by hydrogen bonds among Tyr81, Arg94 and Glu223, whereas EV1 is connected to the CCS¹¹. Unexpectedly, in the dACR iC++, overall extracellular vestibule shapes are more similar to those of the nACR *GtACR1* than the CCR C1C2 (Extended Data Fig. 4)—even though iC++ was designed on the C1C2 backbone—indicating that the changes made to C1C2 when making iC++ may have accessed fundamental features involved in light-activated anion conduction. The realignment of the extracellular vestibule is partly caused by E140S, which breaks the hydrogen bond with Gln95 and allows EV1 to extend to the CCS (Fig. 2a), whereas EV2 is occluded at ECS2 by water- or ion-mediated hydrogen bonding between Arg156 and Arg281 (Fig. 2a).

To analyse stability of constriction site interactions, we performed all-atom molecular dynamics simulations of iC++ and *GtACR1* (Fig. 2b). Whereas overall structures of iC++ and *GtACR1* were stable (Extended Data Fig. 5a), differences were observed. First, interactions at ECS2 of iC++ were weaker than for *GtACR1* (Fig. 2c), in which

interactions among Tyr81, Arg94 and Glu223 (Val146, Arg159 and Arg281 in iC++) were highly stable, and Tyr81–Arg94 remained close to hydrogen-bond distance from Glu223 (Fig. 2d). However in iC++, Arg281 adopted two conformations, such that the Arg156–Arg281 interaction was disrupted and reformed continuously, and water molecules passed from both EV1 and EV2 through ECS2 to the CCS (Fig. 2b, e).

The two conformers of Arg281 exchanged in a chloride-dependent manner (Fig. 2f, left). Simulations started with chloride ions distributed randomly in bulk solvent, yet a chloride ion often spontaneously coordinated between Arg281 and Arg156. In the absence of coordinated chloride, Arg281 exhibited a straight conformation, protruding into the extracellular solvent, but with a coordinated chloride, Arg281 adopted a curved conformation (Fig. 2f). Notably, Arg281 is curved in the crystal structure, and a strong positive peak was detected between Arg156 and Arg281 in $F_o - F_c$ and $2F_o - F_c$ electron-density maps (Extended Data Fig. 5b, c). To assess whether this observed density is a chloride or phosphate in the crystallization condition, we ran refinement in the presence of inorganic phosphate and found a strong negative peak in the $F_o - F_c$ electron-density map (Extended Data Fig. 5d). Furthermore, simulations in the presence of chloride or phosphate revealed that chloride exhibited 10-fold greater occupancy time in the binding site compared to phosphate (Extended Data Fig. 5e). Further studies could evaluate other ions or strongly coordinated water; however, these results suggest that electron density detected in the crystal structure is much more likely to be caused by chloride than phosphate (Fig. 2g). Of note, in the presence of chloride, the curved Arg281 would be more energetically favourable and thereby trapped in the crystal structure. Therefore, one of the functions of Arg281 may be to capture chloride from the extracellular side and facilitate its entry into EV1, as we observed in a simulation run. Consistent with this model, the R281V mutation in iC++ essentially abolished photocurrents while maintaining membrane targeting and expression, demonstrating the functional importance of this structural feature of ECS2 (Fig. 2h, Extended Data Figs. 6–8). Finally, in probing the broad relevance of ECS2 for nACRs, we found that Y81F, R94Q and E223A mutants of *GtACR1* showed an increased slow-closing time (τ_{off2}) (Fig. 2i), illustrating how disruption and reformation of ECS2 hydrogen-bonding interactions may generally regulate channel gating.

Central and intracellular constriction sites

Further insights were gained from analysis of the intracellular side. In the *GtACR1* structure¹¹, Gln46, Glu68, and Asn239 form the CCS, and the intracellular vestibule extends to this CCS (Fig. 3a). By contrast, iC++ revealed two constrictions: the CCS and an intracellular constriction site (ICS; Figs. 2a, 3a). The CCS was mainly formed by Gln129–Gln297 (Glu68–Asn239 in *GtACR1*), and the ICS by Asn122–His173 (Ala61–Leu108 in *GtACR1*). In simulations, interactions at the CCS of *GtACR1* and the ICS of iC++ were highly stable; in *GtACR1*, Gln46–Asn239 maintained hydrogen bonds with Glu68, and in iC++, Asn122 maintained hydrogen bonds with His173 (Fig. 3b, c). By contrast, interaction at the iC++ CCS was unstable; the Gln297 side chain frequently flipped out, disrupting the hydrogen bond with Gln129. Notably, in the flipped conformation, Gln297 formed a new stable van der Waals interaction with Ala126, blocking water permeation (Fig. 3b, d). These results suggest that in *GtACR1* (as proposed^{18,19} for C1C2), the stable hydrogen bond between Glu68 and Asn239 (Glu129 and Asn297 in C1C2) serves to keep the channel closed, but the corresponding interaction in iC++ (Gln129–Gln297) alone does not fully explain inhibition of water or ion flux, as it also requires a stable van der Waals interaction with Ala126 to accommodate the flexibility of Gln297.

To analyse functionality of these residues in more detail, we prepared six CCS mutants, of which five were designed to interconvert structurally corresponding residues of iC++ and *GtACR1* (Fig. 3e, Extended Data Figs. 6–8). Whereas all exhibited robust photocurrents, and the *GtACR1* E68Q mutant (switched to the iC++ residue) maintained high anion selectivity, the corresponding mutant in iC++

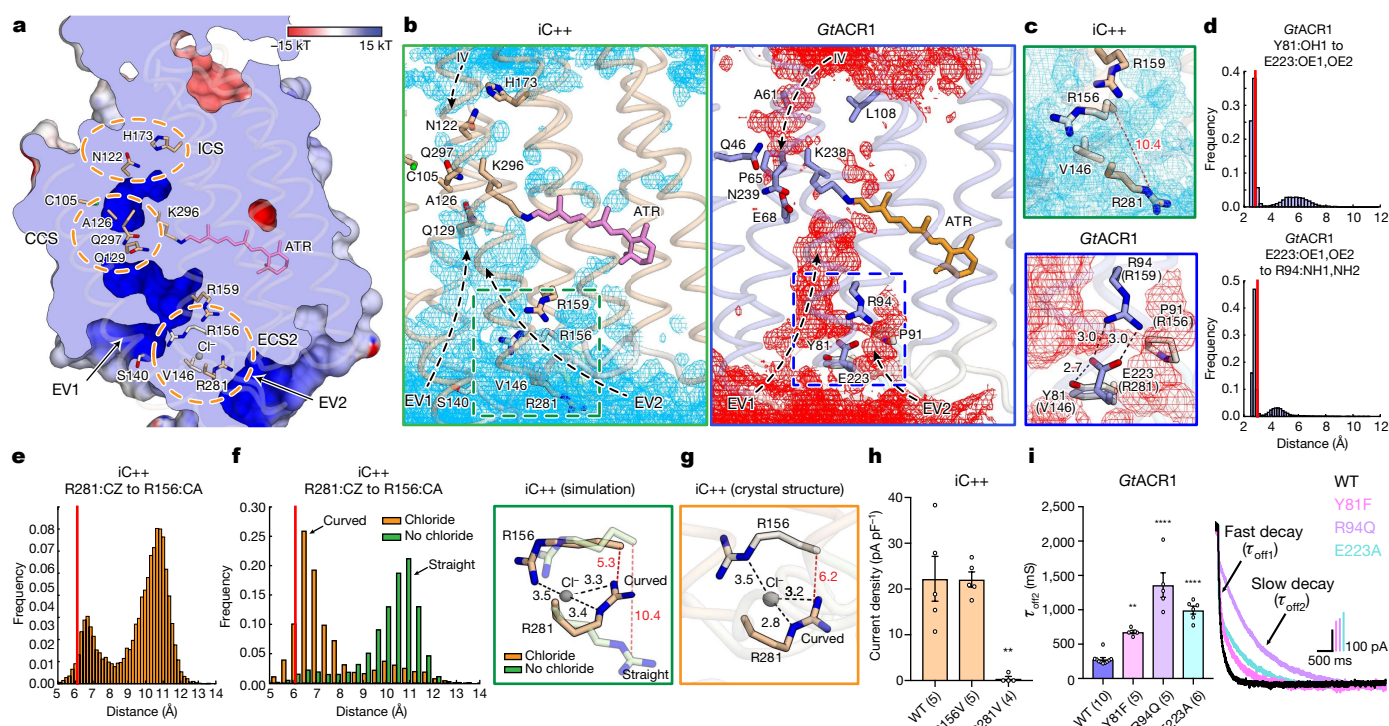


Fig. 2 | Structural and functional characterization of the extracellular constrictions. **a**, Anion-conducting pathway of iC++. Surface is coloured by the electrostatic potential and the orange dashed circles highlight constriction sites. **b**, Simulated water distribution for iC++ (cyan meshes, left) and GtACR1 (red meshes, right). The water density maps are contoured at a probability density of 0.016 molecules per Å³. Dashed boxes, ECS2; arrows, putative water or ion pathways. **c**, Magnified view of the boxed regions from **b**. Black and red dashed lines represent hydrogen-bonding and R156–R281 interactions, respectively, with distances shown

in Å. **d**, **e**, Distance histograms for R281–R156 (**e**), Y81–E223 (**d**, top) and E223–R94 (**d**, bottom). **f**, Distance histograms for R156–R281 (left), and overlaid simulation structures at chloride-binding site with and without chloride (right). **g**, Crystal structure of the chloride-binding site as shown in **f**. **h**, **i**, Effects of mutations in the iC++ chloride-binding site (**h**, $^{**}P = 0.0013$) and GtACR1 ECS2 (**i**, $^{**}P = 0.003$, $^{****}P = 0.0001$). Crystal structure distances are marked as red lines in histograms (**d–f**). Data are mean \pm s.e.m.; one-way ANOVA followed by Dunnett's test. Sample size (number of cells) indicated in parentheses.

(Q129E, switched to the GtACR1 residue), exhibited a depolarized reversal potential (V_{rev}) of -46.3 ± 6.4 mV, indicating reduced anion selectivity. Further, the double CCS mutant in iC++, Q129E/Q297N, exhibited a V_{rev} of -20.93 ± 1.49 mV; therefore, mutating to nACR residues can potentially reduce anion selectivity of iC++ (Fig. 3e). Indeed, under the pore-surface-electrostatic model for channelrhodopsin selectivity^{3,6}, the Q129E mutation (even though it is GtACR1-like) would be expected to reduce the relative level of anion flux (Supplementary Discussion), as observed. Further supporting this surface-electrostatic model rather than the importance of any one specific residue in controlling GtACR1 anion selectivity, whereas Q46A showed slightly depolarized V_{rev} (-54.6 ± 4.85 mV), both GtACR1 to iC++ mutations at the CCS (E68Q and N239Q)—neither of which were predicted to favour cation over anion flux by the pore-surface-electrostatic model—still preserved anion selectivity (Fig. 3e).

In GtACR1, 17 positively charged residues are positioned on the intracellular and extracellular surfaces, with 12 of the 17 located close to the ion-conduction pathway (Fig. 3f). Since iC++ was generated by structure-guided conversion from wild-type cation flux to designed anion flux on the basis of pore-surface electrostatics, a powerful independent confirmation of the fundamental model could potentially be found from structure-guided conversion of GtACR1 in the reverse direction from natural anion flux to designed cation flux. To test this hypothesis, we measured V_{rev} of 20 single and double mutants suggested by analysis of the GtACR1 structure to contribute to pore-surface-electrostatic electropositivity (Fig. 3g, Extended Data Figs. 6–8). Most single mutations in isolation had small effects on V_{rev} , signalling the presence of only moderately increased cation flux (though K188A, K188E and R192E did depolarize V_{rev} to -45.9 ± 4.85 , -47.8 ± 9.15 and -53.7 ± 5.48 mV, respectively). However, the double mutants (Q46A/K188A, K188A/R192A and K188A/R259A)

showed yet further depolarized V_{rev} (-35.2 ± 2.31 , -41.6 ± 4.53 and -39.7 ± 1.84 mV, respectively), confirming that positively charged residues positioned in and near the intracellular pore and vestibules of the ion-conducting pathway contribute to anion selectivity in GtACR1, and demonstrating that CCRs and ACRs can be functionally interconverted, guided only by crystal structures and the pore-surface-electrostatic model.

Structure-guided engineering of FLASH

An emerging theme is that individual residues (depending on local environment) do not necessarily have identical roles in ion selectivity across channelrhodopsin families. Further illustrating this point, interactions at the CCS modulate selectivity much more potently in CCRs than in nACRs (the GtACR1 E68Q and N239Q mutants exhibit V_{rev} similar to the wild type; Fig. 3e). However, residues forming constrictions tend to generally affect closing kinetics. As well as the ECS2 mutant (Fig. 2i; Y81F, R94Q, E223A), all three mutations at the CCS (Q46A, E68Q and N239Q) had powerful effects on off kinetics; most strikingly, τ_{off} of N239Q was more than 20-fold faster than the wild type (Fig. 4a). Because this asparagine on transmembrane helix 7 is highly conserved among nACRs, we analysed this N-to-Q mutation in two other nACRs, GtACR2 and ZipACR²⁰. Although N256Q in ZipACR was nonfunctional, GtACR2(N235Q) exhibited accelerated τ_{off} , confirming that this structure-function relationship can be extended to other nACRs (Extended Data Fig. 6 c,d). Whereas in CCRs many strategies to accelerate kinetics (for higher temporal precision control of spiking) also reduce photocurrent amplitude in cells (owing to reduced operational light sensitivity⁹), the accelerated-decay mutants N239Q in GtACR1 and N235Q in GtACR2 both showed nearly wild-type photocurrent amplitudes (Fig. 4b, Extended Data Figs. 6–8), signalling their promise as inhibitory optogenetic channels.

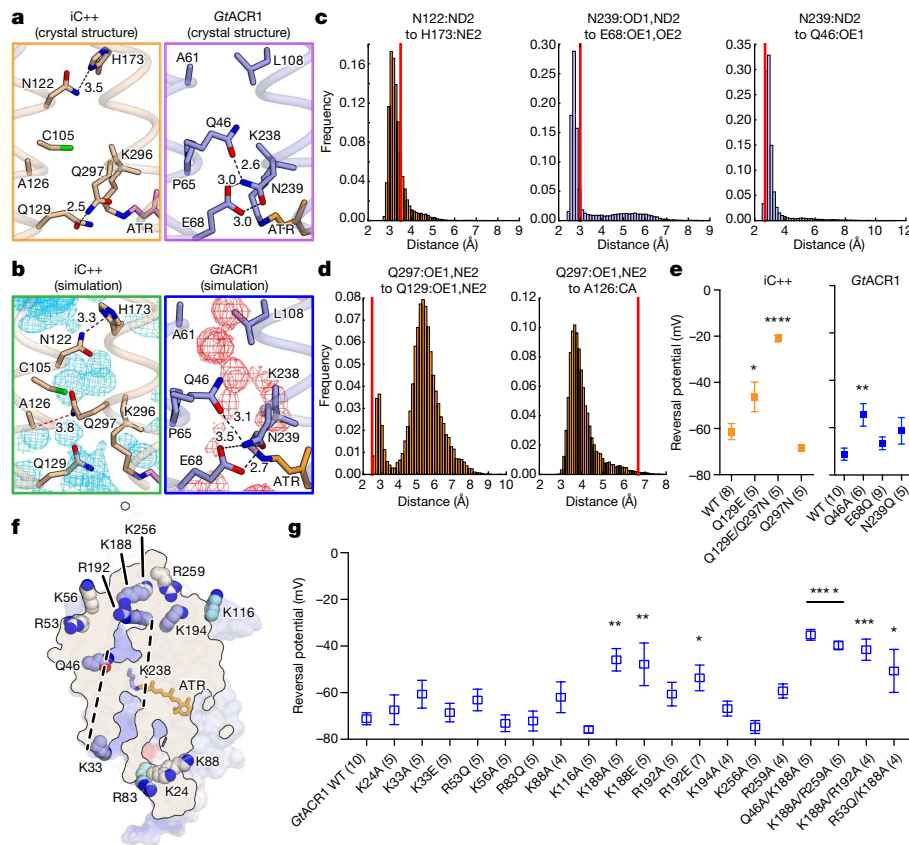


Fig. 3 | Structural basis of ion selectivity in iC++ and GtACR1.

a, Crystal structures of iC++ (left) and GtACR1 (right) constriction sites. Dashed lines represent hydrogen-bonding interactions with distances shown in Å. **b**, Molecular dynamics simulation snapshots of the iC++ (left) and GtACR1 (right) constriction sites. Black and red dashed lines represent hydrogen bonding and van der Waals interactions, respectively. Water density maps are contoured at a probability density of 0.016 molecules per Å³. **c**, **d**, Distance histograms for N122–H173 (left), N239–E68 (middle) and N239–Q46 (right) (**c**), Q297–Q129 (left) and Q297–A126 (right) (**d**) during simulation. **e**, V_{rev} summary for

CCS mutations. (* $P=0.025$, ** $P=0.007$, **** $P=0.0001$, left to right, respectively) **f**, Positively charged residues near the ion-conducting pore of GtACR1. Dashed lines represent the putative ion-conducting path. Residues on helices, β -sheets and loops are coloured blue, cyan and white, respectively. **g**, V_{rev} summary for mutations of the residues shown in **f** (* $P<0.05$, ** $P<0.01$, *** $P=0.0003$, **** $P=0.0001$). Distances in the crystal structure are marked as red lines in histograms (**c**, **d**). Data are mean \pm s.e.m.; one-way ANOVA followed by Dunnett's test. Sample size (number of cells) indicated in parentheses.

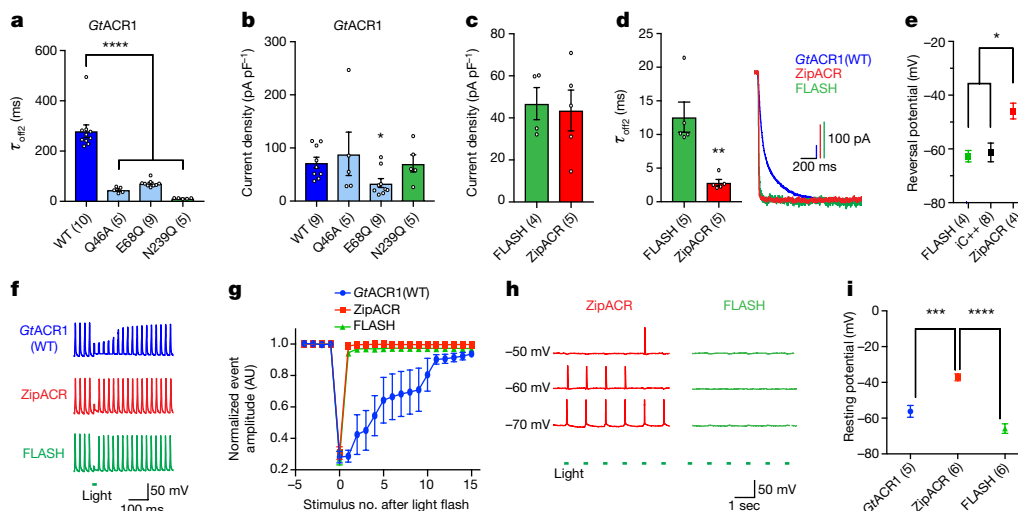


Fig. 4 | Structure-guided engineering of fast and robust channel-based single-spike inhibition for optogenetics. **a**, Summary of off kinetics (**** $P=0.0001$). **b**, Summary of photocurrent densities for the mutants shown in **a** (* $P=0.03$). **c**, Comparison of ACR photocurrent densities ($P=0.8$). **d**, Off kinetics (left) and traces of ACRs at -10 mV holding (right) (** $P=0.003$). **e**, Summary of V_{rev} for different ACRs (* $P<0.03$). **f**, Fast-optogenetic spike suppression by ACRs. **g**, Summary of

depolarizing-event amplitude changes in neurons as shown in **f**. Light delivered at stimulus no. 0. AU, arbitrary units. **h**, Traces illustrating V_{rest} -dependent spiking by ACR stimulation. **i**, Summary of V_{rest} in neurons expressing ACRs. *** $P=0.0004$, **** $P=0.0001$. Data are mean \pm s.e.m.; one-way ANOVA with Dunnett's test (**a**, **i**); one-way ANOVA with Tukey's test (**e**); Kruskal–Wallis test with Dunn's test (**b**); two-tailed t -test (**c**, **d**). Sample size (number of cells) indicated in parentheses.

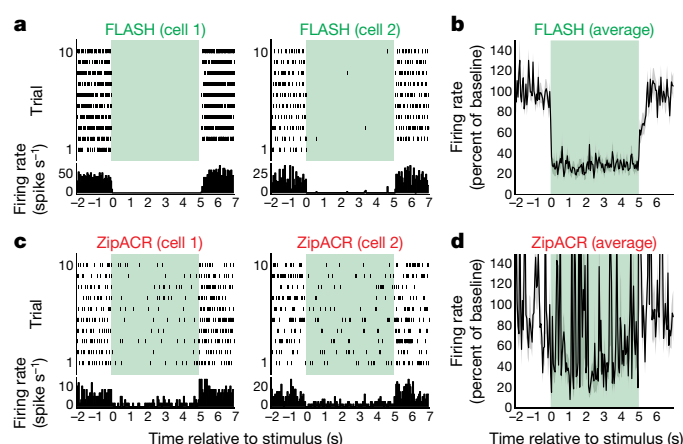


Fig. 5 | Optogenetic application: in vivo electrophysiology in mouse. **a–d** Raster plots, peri-event time histograms (**a**, **c**) and averaged frequency (**b**, **d**) of neuronal spiking in response to light (green area). More cells were modulated by FLASH (24/105) than by ZipACR (9/125) ($n = 230$; one-sided Fisher's test, $P = 0.0011$), where modulated cells exhibited a difference in firing between the 2 s before and the 2 s after illumination ($n = 33$; two-sided Wilcoxon test, $P < 0.05$).

To maximize performance, we further mutated *GtACR1*(N239Q) to R83Q, which we had found to increase photocurrents (Extended Data Figs. 6–8); we designated this *GtACR1*(R83Q/N239Q) double mutant FLASH (fast, light-activated anion-selective rhodopsin). Comparison to ZipACR (which has the fastest reported ACR kinetics)²⁰ in HEK293 cells revealed that FLASH exhibited photocurrents of similar magnitude, moderately slower off kinetics (Fig. 4c, d), and a substantially different V_{rev} , owing to ZipACR's depolarized reversal potential of approximately 15 mV²⁰ (Fig. 4e). We then compared FLASH and ZipACR in cultured neurons; in contrast to *GtACR1*, both FLASH and ZipACR could precisely suppress individual spikes within trains up to 40 Hz (Fig. 4f, g). However, consistent with the depolarized V_{rev} , ZipACR-expressing neurons exhibited occasional light-induced spikes (Fig. 4h) and consistently depolarized V_{rest} (37.17 ± 1.92 mV in the dark; Fig. 4i), relevant considerations for optogenetics.

Comparison of FLASH and ZipACR in vitro and in vivo

We next performed patch-clamp electrophysiology after in vivo expression. Four weeks after injection of opsin-delivering adeno-associated virus vectors (AAVs) into mouse hippocampus, FLASH and ZipACR expressed robustly in acute slices (Extended Data Fig. 9a); however, ZipACR-expressing neurons again displayed depolarized resting potentials (Extended Data Fig. 9b–d). Moreover, consistent with in vitro findings, whereas FLASH and ZipACR exhibited comparable photocurrents (Extended Data Fig. 9e), the ZipACR V_{rev} was approximately 15 mV more depolarized (Extended Data Fig. 9f). In the setting of spikes elicited by 40-Hz current injection (Extended Data Fig. 9g–k), pulsed light inhibited spikes in cells expressing either opsin, but 3 of 8 ZipACR-expressing neurons actually exhibited light-induced spikes, consistent with results from cultured neurons (Fig. 4h, Extended Data Fig. 9k).

Next, we performed extracellular recording in vivo using Neuropixels probes²¹. We injected FLASH or ZipACR AAVs into mouse sensorimotor thalamus, and observed robust expression (Extended Data Fig. 10a, b). Consistent with cultured-neuron and acute-slice patch-clamp data, we observed more efficient net inhibition of spontaneous spiking in FLASH-expressing mice (Fig. 5a–d).

Finally, we expressed FLASH or ZipACR in muscle cells or cholinergic neurons of *Caenorhabditis elegans* (Extended Data Fig. 10c, d) and tested for optogenetic inhibition of behaviour. Swimming of FLASH-expressing but not ZipACR-expressing worms was almost completely eliminated following stimulation with light (Fig. 6), consistent with the mammalian data suggesting that FLASH may represent a suitable ACR for optogenetics.

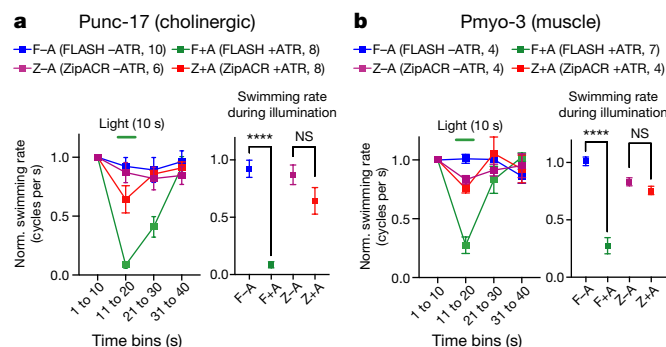


Fig. 6 | Optogenetic application: physiology and behaviour in *C. elegans*. **a**, **b**, Effect of 10-s illumination (from 11 to 20 s) on swimming in animals expressing ACRs in cholinergic neurons (**a**) or muscle (**b**). Data are mean \pm s.e.m. Swimming rate is normalized to the rate during the first 10 s. Data are mean \pm s.e.m.; two-tailed t -test, **** $P < 0.0001$. Sample size (number of animals) indicated in parentheses. NS, not significant.

Discussion

In this work, determination of multiple crystal structures and structure-guided analyses have led to deeper understanding of the channelrhodopsin pore, and to the creation of new tools for optogenetics. Structural and computational analyses of the CCS and ECS2 suggested residues forming constrictions that could be important for kinetics, which led to structure-guided development of FLASH, a fast ACR for high-speed cell-activity suppression, tolerability and efficacy at practical expression levels, hyperpolarized V_{rev} due to high anion selectivity, and fast-off kinetics. Though the more-depolarized V_{rev} of ZipACR was consistent with these analyses, the structural basis for this property had remained unclear. However, our structures inform homology models of ZipACR and FLASH (Extended Data Fig. 9l, m) that suggest that the extracellular surface of FLASH is considerably more electropositive than that of ZipACR, further supporting the importance of residues at and near the vestibule of the ion-conducting pathway in deterring cation flux and controlling V_{rev} of ACRs (Extended Data Fig. 9l, m).

We engineered FLASH by improving the kinetics of *GtACR1*—just one class of channel property that might be advanced using the richness of structural information on *iC++* and *GtACR1*. Structural and functional analyses not only independently confirmed the channelrhodopsin ion-selectivity mechanism, but also suggested a rule for tuning the distinct parameter of photocurrent magnitude. High ion conductance in channelrhodopsins may depend on the overall valence (relative to the conducted species) and distribution of charged residues inside the pore. For example, although both *iC++* and *GtACR1* are highly chloride-selective, *iC++* (with many positively charged residues inside the pore) exhibits lower conductance and reduced photocurrent magnitude in comparison to *GtACR1* (which has its positively charged residues distributed towards the vestibules rather than inside the pore). The latter configuration may reduce the tendency of anions to adhere to the inner surface of the pore (Supplementary Discussion), and thereby, for a given ion selectivity (in this case for anions), powerfully enhance photocurrent magnitude. An inverted valence form of this rule could also apply to CCRs, and straightforward application of the rule could help improve engineered conductances in diverse categories of channelrhodopsins, as well as facilitate sequence-based screening for new channelrhodopsins with high conductance from natural sources.

These structures also point to strategies to modulate ACR absorption spectra, and hence colour selectivity for optogenetics. The crystal structures revealed both similarities and differences in polarity properties of the retinal-binding pocket (RBP) of *iC++* and *GtACR1*. Whereas residues surrounding the polyene of retinal are relatively well conserved, *GtACR1* contains more polar side chains (for example, Cys153 and Ser156) near the β -ionone ring, which presumably contribute to the red-shifted spectrum (peak absorbance wavelength (λ_{max}) of *iC++* and *GtACR1* are 484 and 514 nm, respectively) (Extended Data

Fig. 11a). Supporting the importance of polarity near the β -ionone ring, the S297A mutation in iC++ caused a 10-nm red-shift in spectrum (Extended Data Fig. 11b), similar to C237A, the corresponding mutation in *GtACR1*¹¹. Further, the G220S mutation to the corresponding *GtACR1* residue near the β -ionone in iC++ caused a 6-nm red-shift (Extended Data Fig. 11b). These two mutations are additive; our iC++ (G220S/S295A) double mutant exhibited a 15-nm red-shift compared to wild-type iC++ (Extended Data Fig. 11b), providing proof of principle for structure-guided red-shifting of channelrhodopsin function.

Molecular dynamics simulations also demonstrated that hydration of the RBP differs between iC++ and *GtACR1*; in *GtACR1*, owing to divergence at ICL2¹¹ (Extended Data Fig. 11c), water molecules can penetrate deeply into the dimer interface from the intracellular side and access the RBP from between Cys102 and Ser130 (Cys128 and Asp156 in *Chlamydomonas reinhardtii* ChR2) (Extended Data Fig. 11d). Structure-function insights can also help determine which useful properties are transferable between dACRs and nACRs. Illustrating how key residues may have different roles in distinct local contexts, we found that the CCS was more important for selectivity in iC++, but had a larger influence on kinetics in *GtACR1* (Figs. 3e, 4a).

Additional work is needed to fully understand these and other intricacies, although proof of principle for newly enabled development of next-generation optogenetics has now been established by the bidirectional interconversion of ACRs and CCRs, in generating red-shifted channelrhodopsins, and in the creation of FLASH. More broadly, this study provides a framework for understanding light-induced ion conduction, by providing insights into principles governing selectivity, photocurrent magnitude, kinetics and action spectrum relevant to all channelrhodopsins.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0504-5>.

Received: 23 February 2018; Accepted: 13 August 2018;

Published online 29 August 2018.

- Deisseroth, K. & Hegemann, P. The form and function of channelrhodopsin. *Science* **357**, eaan5544 (2017).
- Kato, H. E. et al. Crystal structure of the channelrhodopsin light-gated cation channel. *Nature* **482**, 369–374 (2012).
- Berndt, A., Lee, S. Y., Ramakrishnan, C. & Deisseroth, K. Structure-guided transformation of channelrhodopsin into a light-activated chloride channel. *Science* **344**, 420–424 (2014).
- Wietek, J. et al. Conversion of channelrhodopsin into a light-gated chloride channel. *Science* **344**, 409–412 (2014).
- Wietek, J. et al. An improved chloride-conducting channelrhodopsin for light-induced inhibition of neuronal activity in vivo. *Sci. Rep.* **5**, 14807 (2015).
- Berndt, A. et al. Structural foundations of optogenetics: Determinants of channelrhodopsin ion selectivity. *Proc. Natl Acad. Sci. USA* **113**, 822–829 (2016).
- Govorunova, E. G., Sineshchekov, O. A., Janz, R., Liu, X. & Spudich, J. L. Natural light-gated anion channels: a family of microbial rhodopsins for advanced optogenetics. *Science* **349**, 647–650 (2015).
- Kim, C. K., Adhikari, A. & Deisseroth, K. Integration of optogenetics with complementary methodologies in systems neuroscience. *Nat. Rev. Neurosci.* **18**, 222–235 (2017).
- Gunaydin, L. A. et al. Ultrafast optogenetic control. *Nat. Neurosci.* **13**, 387–392 (2010).
- Yizhar, O. et al. Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* **477**, 171–178 (2011).

- Kim, Y.S. et al. Crystal structure of a natural anion-conducting channelrhodopsin, *GtACR1*. *Nature* <https://doi.org/10.1038/s41586-018-0511-6> (2018).
- Deisseroth, K. Optogenetics: 10 years of microbial opsins in neuroscience. *Nat. Neurosci.* **18**, 1213–1225 (2015).
- Allen, W. E. et al. Thirst-associated preoptic neurons encode an aversive motivational drive. *Science* **357**, 1149–1155 (2017).
- Chung, S. et al. Identification of preoptic sleep neurons using retrograde labelling and gene profiling. *Nature* **545**, 477–481 (2017).
- Mamad, O. et al. Place field assembly distribution encodes preferred locations. *PLoS Biol.* **15**, e2002365 (2017).
- Selimbeyoglu, A. et al. Modulation of prefrontal cortex excitation/inhibition balance rescues social behavior in *CNTNAP2*-deficient mice. *Sci. Transl. Med.* **9**, eaah6733 (2017).
- Sineshchekov, O. A., Li, H., Govorunova, E. G. & Spudich, J. L. Photochemical reaction cycle transitions during anion channelrhodopsin gating. *Proc. Natl Acad. Sci. USA* **113**, E1993–E2000 (2016).
- Takemoto, M. et al. Molecular dynamics of channelrhodopsin at the early stages of channel opening. *PLoS One* **10**, e0131094 (2015).
- VanGordon, M. R., Gyawali, G., Rick, S. W. & Rempe, S. B. Atomistic study of intramolecular interactions in the closed-state channelrhodopsin chimera, C1C2. *Biophys. J.* **112**, 943–952 (2017).
- Govorunova, E. G. et al. The expanding family of natural anion channelrhodopsins reveals large variations in kinetics, conductance, and spectral sensitivity. *Sci. Rep.* **7**, 43358 (2017).
- Jun, J. J. et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017).

Acknowledgements We thank C. Lee, M. Lo, K. Geiselhart and M. Lima for technical support; K.K. Kumar, N.R. Latorraca, M. Inoue and K. Katayama for comments on the manuscript; E.E. Steinberg, M.A. Wright, and R.C. Malenka for inputs on mouse experiments; K. Hirata, M. Yamamoto and other staffs at BL32XU of SPring-8, and the staff at BL23ID-B/23ID-D of APS for assistance in data collection. We acknowledge support by JST PRESTO (JPMJPR1782 for H.E.K., JPMJPR15P2 for K.I.), Stanford Bio-X and the Kwanjeong Foundation (Y.S.K.), German Academic Exchange Service (D.H.); MEXT (17H03007 for K.I. and 25104009/15H02391 for H.K.), JST CREST (JPMJCR1753, H.K.), the US Department of Energy, Scientific Discovery through Advanced Computing (SciDAC) program (R.O.D.) and Mathers Charitable Foundation (B.K.K.). The project was supported by a grant for channelrhodopsin crystal structure determination from the NIMH (R01MH075957 to K.D.).

Reviewer information *Nature* thanks P. Scheerer, L. Tian and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions H.E.K. and Y.S.K. contributed equally and either has the right to list himself first in bibliographic documents. H.E.K. and Y.S.K. expressed, purified, and crystallized iC++, harvested crystals and collected diffraction data. H.E.K. and K.Y. solved the structures and analysed electron densities. H.E.K. and Y.S.K. measured UV-vis spectra of iC++ mutants. Y.S.K. and L.E.F. performed electrophysiology. J.M.P. performed and analysed molecular dynamics simulations under guidance of R.O.D. W.E.A. and K.E.E. conducted in vivo recording. S.I. and K.I. measured UV-vis spectra and performed flash photolysis under guidance of H.K. C.Ra. and K.E.E. performed neuron cultures and molecular cloning. C.Ri. and Y.S.K. performed *C. elegans* experiments under guidance of K.S. D.H. performed crystallography and S.Y.L. and A.B. performed electrophysiology at an early stage. K.D. initiated and supervised this channelrhodopsin structure-function project. H.E.K., Y.S.K., B.K.K. and K.D. planned and guided the work, and interpreted the data. H.E.K., Y.S.K. and K.D. prepared the manuscript and wrote the paper with input from all the authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0504-5>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0504-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to H.E.K. or K.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample sizes were determined based on prior literature and best practices in the field; no statistical methods were used to predetermine sample size.

Cloning, protein expression, and purification. The previously described iC++ construct⁶ was used for crystallography after introducing a point mutation (N61Q) to remove an *N*-glycosylation site. A FLAG tag followed by the 3C protease cleavage site was added to the N terminus and an enhanced GFP (EGFP) with a His₁₀ tag and the 3C site was attached to the C terminus. The finalized iC++ crystallization construct was expressed in Sf9 cells (Expression Systems, identified by the vendor, not tested for mycoplasma contamination) using the BestBac (Expression Systems) baculovirus system. Cell cultures were grown to a density of 4×10^6 cells ml⁻¹, infected with iC++ baculovirus, and were shaken at 27 °C for 18 h. Then, 20 μM all-*trans* retinal (ATR) was added to the culture and incubation continued for a further 54 h, and cell pellets were harvested and stored at -80 °C. To purify iC++, the pellets were lysed with a hypotonic lysis buffer (20 mM HEPES pH 7.5, 1 mM EDTA and protease inhibitors). The cell debris was then homogenized with a glass dounce homogenizer in a solubilization buffer (1% *n*-dodecyl-β-D-maltopyranoside (DDM), 0.06% cholesteryl hemisuccinate tris salt (CHS), 20 mM HEPES pH 7.5, 500 mM NaCl, 20% glycerol, 10 mM imidazole, and protease inhibitors) and solubilized for 2 h at 4 °C. The insoluble cell debris was removed by centrifugation (18,000 r.p.m., 25 min), and the supernatant was mixed with the Ni-NTA agarose resin (Qiagen) for 2 h at 4 °C. The Ni-NTA resin was collected into a glass chromatography column, washed with 20 column volumes of a wash buffer (0.05% DDM, 0.01% CHS, 20 mM HEPES pH 7.5, 500 mM NaCl, 20% glycerol and 20 mM imidazole) and was eluted in a wash buffer supplemented with 250 mM imidazole. The Ni-NTA eluent was then supplemented with 2 mM CaCl₂ and was loaded over anti-FLAG M1 resin over 1 h. The protein was then washed with a Flag wash buffer (0.05% DDM, 0.01% CHS, 20 mM HEPES pH 7.5, 300 mM NaCl, 5% glycerol and 2 mM CaCl₂) and eluted with a Flag elution buffer (0.05% DDM, 0.01% CHS, 20 mM HEPES pH 7.5, 300 mM NaCl, 5% glycerol, 0.2 mg ml⁻¹ Flag peptide and 3 mM EDTA). Following the cleavage of Flag tag and EGFP-His₁₀ by His-tagged 3C protease, the sample was reloaded onto the Ni-NTA column to capture the cleaved EGFP-His₁₀. The flow-through containing iC++ N61Q was collected, concentrated and purified through gel-filtration chromatography in a final buffer (100 mM NaCl, 20 mM HEPES pH 7.5, 0.05% DDM and 0.01% CHS). Peak fractions were pooled and concentrated to 15 mg ml⁻¹.

Crystallization. Purified iC++ protein was crystallized using LCP method as described previously². Best crystals were obtained in 28% PEG 350MME, 100 mM Tris pH 8.5, 100 mM ammonium phosphate dibasic, and 7–8% formamide for pH 8.5, and, 30% PEG 350MME, 100 mM sodium phosphate pH 6.5, and 200 mM sodium malonate for pH 6.5. All crystals were harvested using micromesh loops (MiTeGen), and were flash-cooled in liquid nitrogen without any additional cryoprotection for data collection.

Data collection and structure determination. X-ray diffraction data of iC++ at pH 8.5 and pH 6.5 were collected at Advanced Photon Source GM/CA-CAT beamline 23ID-B and 23ID-D, and SPring-8 BL32XU, respectively. Small wedge data, each consisting of 5–20°, were collected from single crystals at wavelengths of 1.03 and 1.00 Å, respectively. Collected datasets (58 datasets for pH 8.5 and 442 datasets for pH 6.5) were processed automatically using KAMO²² (<https://github.com/keitaroyam/yamtbx/blob/master/doc/kamo-en.md>). Each dataset was indexed and integrated using XDS²³, and classified using the correlation coefficients of the normalized structure amplitudes between datasets. A total of 16 (pH 8.5) and 27 (pH 6.5) datasets in best cluster were scaled and merged using XSCALE. The best clusters were found from those giving lower inner-shell R_{meas} and higher outer-shell $CC_{1/2}$ values. The structure was determined by molecular replacement with the program Phaser²⁴, using the cation channelrhodopsin chimaera C1C2 (PDB ID: 3UG9) as the search model. The resultant structure was iteratively refined in Refmac²⁵ and Phenix²⁶, and manually rebuilt in Coot²⁷. The final models of iC++ at pH 6.5 and pH 8.5 contained 97.1, 2.9 and 0.0%, and 96.0, 4.0, and 0.0% in the favoured, allowed, and outlier regions of the Ramachandran plot, respectively. Final refinement statistics are summarized in Extended Data Fig. 1e. All molecular graphics figures were prepared with Cuemol (<http://www.cuemol.org>).

Hippocampal neuron and HEK293 cell electrophysiology. Recordings in hippocampal cultured neurons were performed 4–6 days after transfection in Tyrode's solution: 150 mM NaCl, 4 mM KCl, 2 mM CaCl₂, 2 mM MgCl₂, 10 mM glucose and 10 mM HEPES-NaOH pH 7.4. Intracellular solution contained 140 mM K-gluconate, 10 mM HEPES-KOH pH 7.2, 10 mM EGTA and 2 mM MgCl₂. Signals were amplified and digitized using the AxoPatch200B and DigiData1400 (Molecular Devices). The Spectra X Light engine (Lumencor) served as a light source, and 470/15 and 513/15 filters were used for blue and green light respectively, followed by coupling into a Leica DM LFSA microscope. Patch pipettes (3–6 MΩ) were pulled using a P2000 micropipette puller (Sutter Instruments). HEK293 cells were cultured as previously described⁶. Cells were

transfected using Lipofectamine 2000 (Life Technologies). Recordings in HEK293 cells (Thermo Fisher, identified by the vendor, not tested for mycoplasma contamination) were performed 24–48 h after transfection in Tyrode's solution and intracellular solution as described above.

Voltage clamp recordings in neurons and HEK cells were performed in the presence and absence of bath-applied tetrodotoxin (TTX, 1 μM; Tocris), respectively. Although previous work (while reporting the same depolarizing V_{rev} for ZipACR that we observed) had not also reported an effect on neuronal V_{rest} ²⁰, an unusual promoter was used in that work²⁰ (from the ubiquitin gene, not typically used for optogenetics owing to relative weakness of expression). Here a stronger promoter (widely used in optogenetics, from CaMKIIα) was used in vitro. Likewise in vivo expression in mammalian brain tissue, to match typical conditions for practical optogenetics we constructed an adeno-associated viral vector (AAV) carrying the genes encoding FLASH or ZipACR fused to enhanced YFP (eYFP) and a membrane-trafficking sequence (TS)²⁸, under the control of the CaMKIIα promoter. For reversal potential measurements in vitro, cells were held at resting potentials from -95 mV to +15 mV in steps of 10 mV, with 1.0 mW mm⁻² light (470 nm and 513 nm for iC++ and GtACR1, respectively) delivered for 1.5 s. Channel kinetics and photocurrent amplitude were measured at -10 mV and -70 mV holdings, respectively. The same light stimulus used for reversal potential measurement. Liquid junction potentials (LJPs) were corrected using the Clampex built-in LJP calculator as previously described⁶. Peak photocurrent amplitudes were normalized to each cell's membrane capacitance, which was calculated from the Clampex built-in membrane test. Current clamp measurements in neurons were conducted in the presence of glutamatergic synaptic blockers: 6-cyano-7-nitroquinoxaline-2,3-dione (CNQX; 10 μM, Tocris) for AMPA receptors and D(-)-2-amino-5-phosphonopivalic acid (APV; 25 μM, Tocris) for NMDA receptors. For photoinduced spike suppression, neurons were stimulated by injection of 5 ms of 500 pA current at 40 Hz, and 10-ms light pulses of 513 nm light at 1.0 mW mm⁻² were applied during the recording. For photoinduction of spikes in ZipACR and FLASH expressing neurons, 10-ms light pulses of 513 nm light at 1.0 mW/mm² were applied at 10 Hz. Analyses of electrophysiology results were performed as previously described³. Statistical analysis was performed with a two-tailed *t*-test or a one-way ANOVA, and a Kruskal–Wallis test for non-parametric data, using Prism 7 (GraphPad) software.

pH titration. To investigate pH-dependency of absorption spectrum of iC++, ~3 μM protein was solubilized in 6-mix buffer (150 mM NaCl, 4 mM KCl, 10 mM glucose, 2 mM CaCl₂, 2 mM MgCl₂, 6-mix buffer (10 mM citrate, 10 mM MES, 10 mM HEPES, 10 mM MOPS, 10 mM CHES, 10 mM CAPS) pH 8.5, 0.05% DDM, 0.01% CHS). Then, pH was decreased by addition of concentrated HCl to pH 3.0. Absorption spectra measured with a UV-visible spectrometer (V-2400PC, SHIMADZU) for every ~0.5 pH unit change.

Laser flash photolysis. Transient absorption changes after photo-excitation were investigated by laser flash photolysis. iC++ was solubilized in 150 mM NaCl, 10 mM HEPES (pH 7.4, NaOH), 4 mM KCl, 10 mM glucose, 2 mM CaCl₂, 2 mM MgCl₂, 0.05% DDM, 0.01% CHS. Optical density of the suspension was adjusted to be 0.8–0.9 by dilution (~8.5 μM protein concentration). The sample solution was illuminated with a second harmonic generation by a nano-second pulsed Nd³⁺:YAG laser (λ = 532 nm, INDI40, Spectra-Physics) with the pulse energy of 3.8 mJ cm⁻² pulse. The transient absorption spectra of iC++ after laser excitation were obtained by measuring the intensity of white light passed through the sample before and after laser excitation at λ = 300–700 nm with an ICCD linear array detector (C8808-01, Hamamatsu). To increase signal-to-noise ratio, 90 identical spectra were averaged and singular-value-decomposition (SVD) analysis was applied. Time evolution of transient absorption change at specific wavelength after photo-excitation was measured by monitoring the change in intensity of monochromated output of a Xe arc lamp (L9289-01, Hamamatsu Photonics) passed through the sample solution by a photomultiplier tube (R10699, Hamamatsu Photonics) equipped with a notch filter (532 nm, bandwidth = 17 nm, Semrock) to avoid unnecessary detection of scattered pump pulse. The signals were monitored and stored by a digital oscilloscope (DPO7104, Tektronix). To increase signal-to-noise ratio, 50–100 identical signals were averaged.

Measurement of UV absorption spectra. Protein absorbance spectra were measured with an Infinite M1000 microplate reader (Tecan Systems Inc.) using 96-well plates (ThermoFisher Scientific). The GtACR1 samples were suspended in a buffer containing 100 mM NaCl, 0.05% DDM, 0.01% CHS, and 20 mM sodium citrate, sodium acetate, sodium cacodylate, HEPES, Tris, CAPSO, or CAPS. The pH was adjusted from pH 4 to pH 10 by the addition of NaOH or HCl.

Molecular dynamics simulations. System setup for molecular dynamics simulations. Molecular dynamics simulations were performed for iC++ and GtACR1. Simulations of GtACR1 were initiated from coordinates closely resembling chains C and D of PDB structure 6CSM (that is, from a slightly earlier refinement). Simulations of iC++ were initiated from coordinates closely resembling the structure described in this manuscript. For both GtACR1 and iC++, additional

short simulations of the final refined structures were performed to ensure that there were no significant differences in the dynamics. All proteins were simulated as dimers. We performed five replicates of simulations for all three systems. For each replicate, initial atom velocities were assigned randomly and independently. Prime (Schrödinger) was used to model missing side chains and loops, and neutral acetyl and methylamide groups were added to cap protein termini. Titratable residues remained in their dominant protonation state at pH 7, as determined using PropKa, except that in *GtACR1*, Asp234 and Glu68 were protonated as is consistent with the spectroscopy data in the accompanying paper¹¹. Dowser software was used to add waters to cavities within the protein structure²⁹. The prepared protein structures were aligned to the Orientation of Proteins in Membranes (OPM) structure for PDB 3UG9³⁰. The aligned structures were then inserted into a pre-equilibrated palmitoyl-oleoyl-phosphatidylcholine (POPC) bilayer using Dabble, a simulation preparation software³¹. Sodium and chloride ions were added to neutralize each system at a concentration of approximately 150 mM. Bilayer dimensions were chosen to maintain at least a 30 Å buffer between protein images in the *x-y* plane and a 20 Å buffer between protein images in the *z* direction. Final system dimensions were approximately 93 × 93 × 115 Å³. Simulation times in ns for each replicate are 1749, 1718, 1741, 1673, 1803 for iC++ and 2296, 2285, 2247, 1128 and 2194 for *GtACR1*.

Molecular dynamics simulation protocols. We used the CHARMM36m force field for proteins, lipids, and ions, and the TIP3P model for waters^{32–36}. Parameters for retinal were obtained through personal communication with Scott Feller³⁷. We performed simulations using the Compute Unified Device Architecture (CUDA) version of particle-mesh Ewald molecular dynamics (PMEMD) in AMBER on one or two graphics processing units (GPUs)³⁸. Simulations were performed using the AMBER16 software³⁹. Three rounds of minimization were performed, each consisting of 500 iterations of steepest descent minimization, followed by 500 iterations of conjugate gradient descent minimization, with harmonic restraints of 10.0, 5.0, and 1.0 kcal·mol^{−1}·Å^{−2} placed on the protein and lipids in respective rounds. Systems were heated from 0K to 100K in the NVT ensemble over 12.5 ps and then from 100K to 310K in the NPT ensemble over 125 ps, using 10.0 kcal·mol^{−1}·Å^{−2} harmonic restraints applied to lipid and protein heavy atoms. Systems were then equilibrated at 310K in the NPT ensemble at 1 bar, with harmonic restraints on all protein heavy atoms tapered off by 1.0 kcal·mol^{−1}·Å^{−2} starting at 5.0 kcal·mol^{−1}·Å^{−2} in a stepwise fashion every 2 ns for 10 ns and then by 0.1 kcal·mol^{−1}·Å^{−2} in a stepwise fashion every 2 ns for 20 ns. Production simulations were performed in the NPT ensemble at 310K and 1 bar, using a Langevin thermostat for temperature coupling and a Monte Carlo barostat for pressure coupling. These simulations used a 4-fs time step with hydrogen mass repartitioning⁴⁰. Bond lengths to hydrogen atoms were constrained using SHAKE. Simulations used periodic boundary conditions. Non-bonded interactions were cut off at 9.0 Å, and long-range electrostatic interactions were computed using particle-mesh Ewald (PME) with an Ewald coefficient of approximately 0.31 Å and an interpolation order of 4. The FFT grid size was chosen such that the width of a grid cell was approximately 1 Å.

Analysis protocols for molecular dynamics simulation. Trajectory snapshots were saved every 200 ps during production simulations. The AmberTools15 CPPTRAJ package was used to reimage and centre trajectories⁴¹. Simulations were visualized and analysed using Visual Molecular Dynamics (VMD)⁴². Water density maps were generated using an in-house Python script. Frames representing every 1 ns of simulation, excluding the first 500 ns, were used as input. All r.m.s.d. and atom distance plots were produced using in-house scripts with VMD's python modules. All histograms of atom-to-atom distances were calculated for frames representing every 1 ns of simulation, excluding the first 100 ns, from all relevant simulations. **Determining the relative binding affinities of chloride and phosphate.** To determine the relative binding affinity of phosphate and chloride, we built two systems: one that contained 150 mM chloride and another in which all chlorides were removed and half were replaced with HPO₄^{2−}, the most relevant species of inorganic phosphate at physiological pH (especially for one that would be coordinated with positively charged residues). Otherwise the systems were identical. These systems were prepared following the protocol described above, except that the structure of iC++ to be deposited in the PDB was used instead of an earlier refinement. Partial charges and bonded parameter terms for phosphate were obtained through CGenFF⁴³, but we increased the Lennard Jones radii of the phosphate oxygens by 0.5 Å to prevent the aggregation of phosphates and sodium into a crystal, following precedent in the AMBER literature⁴⁴. For the chloride system, we carried out 10 simulations of lengths 376, 982, 976, 980, 994, 986, 441, 991, 488, and 401 ns. For the phosphate system, we carried out 7 simulations of length 917, 890, 980, 812, 916, 944, and 970 ns. In our analysis, we aligned the protein on its transmembrane helices, and considered the binding pocket to be occupied if the centre of a chloride or phosphate is present within 4 Å of the modelled position (corresponding to the sphere in Extended Data Fig. 5e). We considered the binding site in each monomer independently.

Stereotactic surgeries. All mouse experiments conformed to guidelines established by the National Institutes of Health and were conducted under protocols approved by the Stanford Administrative Panel on Laboratory Animal Care (protocols #11414). All mice were group-housed in a light-regulated colony room (lights on at 07:00, off at 19:00), with food and water available ad libitum. Wild-type male and female C57BL/67 mice were obtained from Jackson Laboratory.

All stereotactic surgeries were performed with mice under isoflurane anaesthesia (4% initially, maintained at 2–3%) with regular monitoring for stable respiratory rate and absent tail pinch response. The scalp was shaved and mice were placed in the stereotactic apparatus and a heating pad was used to prevent hypothermia. All coordinates are measured in millimetres from bregma as defined⁴⁵.

A midline incision was made to expose the skull and small craniotomies were made above the injection sites using a Meisinger Carbide Burr size ¼. All virus dilutions were performed in ice-cold PBS and all viruses were produced at the Stanford Gene and Viral Vector Core. Virus injections were delivered with a 10-µl syringe (World Precision Instruments) and 33-gauge bevelled needle (World Precision Instruments), injected at 100 nl min^{−1} using an injection pump (World Precision Instruments). Following injection, the injection needle was held at the injection site for 10 min then slowly withdrawn. Mice were administered 0.5–1.0 mg kg^{−1} subcutaneous buprenorphine-SR (ZooPharma) approximately 30 min before the end of the surgery for post-operative pain management.

For acute slice recordings, mice were injected bilaterally in the CA1 (ML: ±1.6, AP: −2.5, DV: −1.5) and DG (ML: ±1.6, AP: −2.5, DV: −2.1) regions of the hippocampus. At each coordinate, 1,000 nl of virus was delivered at a titer of 6.75 × 10¹² genome copies (gc)/ml. Four C57BL/67 mice were injected with AAV8-CaMKIIa-ZIPACR-eYFP, and four C57BL/67 mice were injected with AAV8-CaMKIIa-FLASH-eYFP.

For optrode recordings, mice were injected unilaterally in thalamus (ML: +1.25, AP: −1.3, DV: −3.5 and ML: +1.4, AP: −2.4, DV: −3.3) and hippocampus (ML: +1.25, AP: −1.3, DV: −1.75 and ML: +1.4, AP: −2.4, DV: −1.75). At each thalamus site, 1,000 nl of virus was delivered and at each hippocampus site 500 nl of virus was delivered. Mice received either AAV8-CaMKIIa-ZIPACR-eYFP or AAV8-CaMKIIa-FLASH-eYFP diluted to a concentration of 6.74 × 10¹² gc/ml. Mice were implanted with a 200-µm core diameter, 0.39-NA fibre (Thorlabs; CFMLC12L05) dorsal to the thalamus at a 30-degree angle (ML: +3.5, AP: −1.9, DV: −2.4). Mice were affixed with a headbar, implanted with a reference electrode, and the skull encased in a clear, UV-curing resin. Two weeks post-virus injection, an acute craniotomy was drilled over the thalamus and the acute recording electrode implanted. We recommend for safe and effective expression of ACRs (including FLASH) a viral titer of <2 × 10¹² gc/ml, and expression time of 2–4 weeks post injection.

Slice electrophysiology. Acute slice recordings were performed 4–5 weeks after virus injection. Coronal slices 300 µm in thickness were prepared after intracardial perfusion with ice-cold *N*-methyl-D-glucamine (NMDG) containing cutting solution: 93mM NMDG, 2.5 mM KCl, 25 mM glucose, 1.2 mM NaH₂PO₄, 10 mM MgSO₄, 0.5 mM CaCl₂, 30 mM NaHCO₃, 5 mM Na ascorbate, 3 mM Na pyruvate, 2 mM thiourea and 20 mM HEPES pH 7.3–7.4. Slices were incubated for 12 min at 32–34 °C, and then were transported to room temperature oxygenated artificial cerebrospinal fluid (ACSF) solution: 124 mM NaCl, 2.5 mM KCl, 24 mM NaHCO₃, 2 mM CaCl₂, 2 mM MgSO₄, 1.2 mM NaH₂PO₄, 12.5mM glucose and 5mM HEPES pH 7.3–7.4. The ACSF also contained synaptic transmission blockers 25 µM APV and 10 µM CNQX for recordings. Recording patch pipettes contained the following intracellular solution: 140 mM K-gluconate, 10 mM HEPES pH 7.2, 10 mM EGTA and 2 mM MgCl₂.

Cell input resistance and capacitance were calculated after a −10-mV voltage step. Then, the mode was switched to current clamp to determine the resting potential of each cell and the holding current to keep the cell's membrane potential at −70 mV, followed by stepwise current injection to determine the threshold current for spike generation. For all experiments, 513-nm light at 1 mW mm^{−2} was used. For single-spike suppression, neurons were stimulated by 40-Hz, 5-ms pulsed current injection and 10-ms light was delivered. For multiple spike suppression, neurons were illuminated with 1 s of 10-ms pulsed light at 40 Hz, to match the electrical stimulation. For photo-induction of spikes in ZipACR and FLASH expressing neurons, 30-ms light pulses were applied at 10 Hz without current injection. Then, the mode was switched back to voltage clamp and *V_m* was held at −10 mV, whereupon 1-s light was delivered for measuring photocurrent magnitude.

In vivo electrode recording. Extracellular electrophysiological data were recorded using a Neuropixels Phase 3 Option 3 probe, referenced using a Pt–Ir wire inserted into the anterior cortex. Recordings were performed targeting the centre of the virally injected area in the thalamus. The electrode was coated with DiI (ThermoFisher) to reconstruct the tract. 384 channels were acquired at 30 kHz. Following common average referencing, well-isolated single units were identified using KiloSort and Phy. ZipACR and FLASH were activated using continuous

532-nm illumination (for 5 s, at 5 mW at the fibre tip) delivered via an optical fibre (200 μ m) placed at the surface of the surgery area. The optical fibre was connected to a patch cable coupled to a green Laser Diode Fibre Light Source (Doric lenses). Changes in firing rate relative to the light onset were analysed using custom Python scripts.

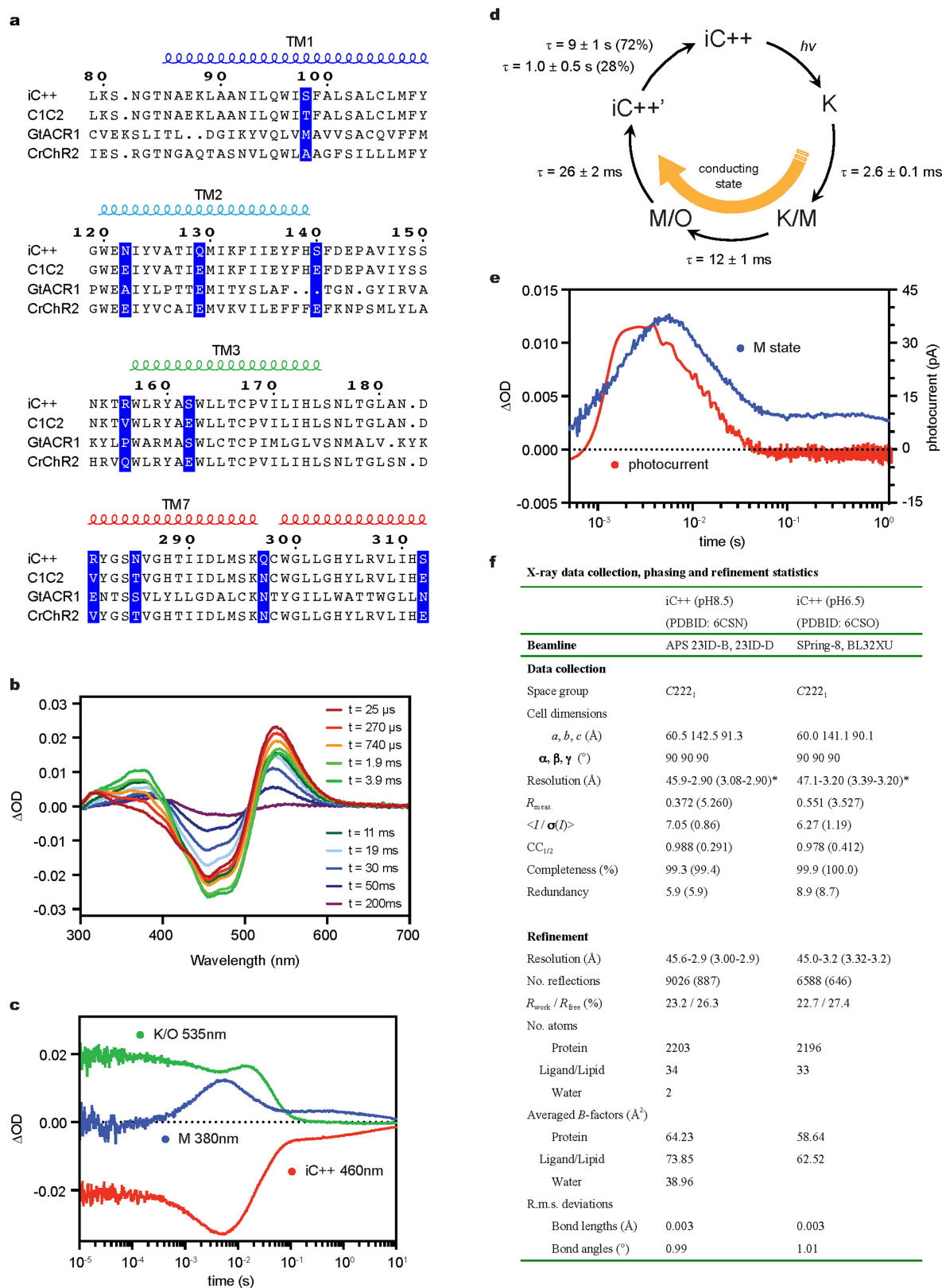
Histological procedures. Mice were deeply anesthetized with sodium pentobarbital and transcardially perfused with 4% paraformaldehyde. Brains were removed, post-fixed for 24 h and sectioned at 100 μ m at room temperature on a vibratome. Slides coverslipped with Vectashield HardSet AntiFade mounting media with DAPI (Vector Laboratories). Images were taken using a Leica DM6000 B confocal laser microscope.

C. elegans experiments. Genes encoding opsin-GFP were placed under the muscle-specific *myo-3* promoter or under the cholinergic motor neuron specific *unc-17* promoter, and lines carrying extrachromosomal arrays were generated using *unc-119* rescue. *C. elegans* were grown in the presence or the absence of ATR, as described previously⁴⁶, but with 20 μ M ATR. For optogenetic manipulation of swimming, a 10-s epoch of the animal swimming was first recorded as a baseline, followed by 10 s of illumination and then 20 s without light. For light delivery, Zeiss FluoArc 120W/45C VIS mercury lamp with a Zeiss Axioplan 2 microscope using a 546/12 excitation filter (estimated power = 43.1 mW cm⁻²) was used. All experiments were randomized and experimenters were blinded to allocation during experiment and outcome assessment.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

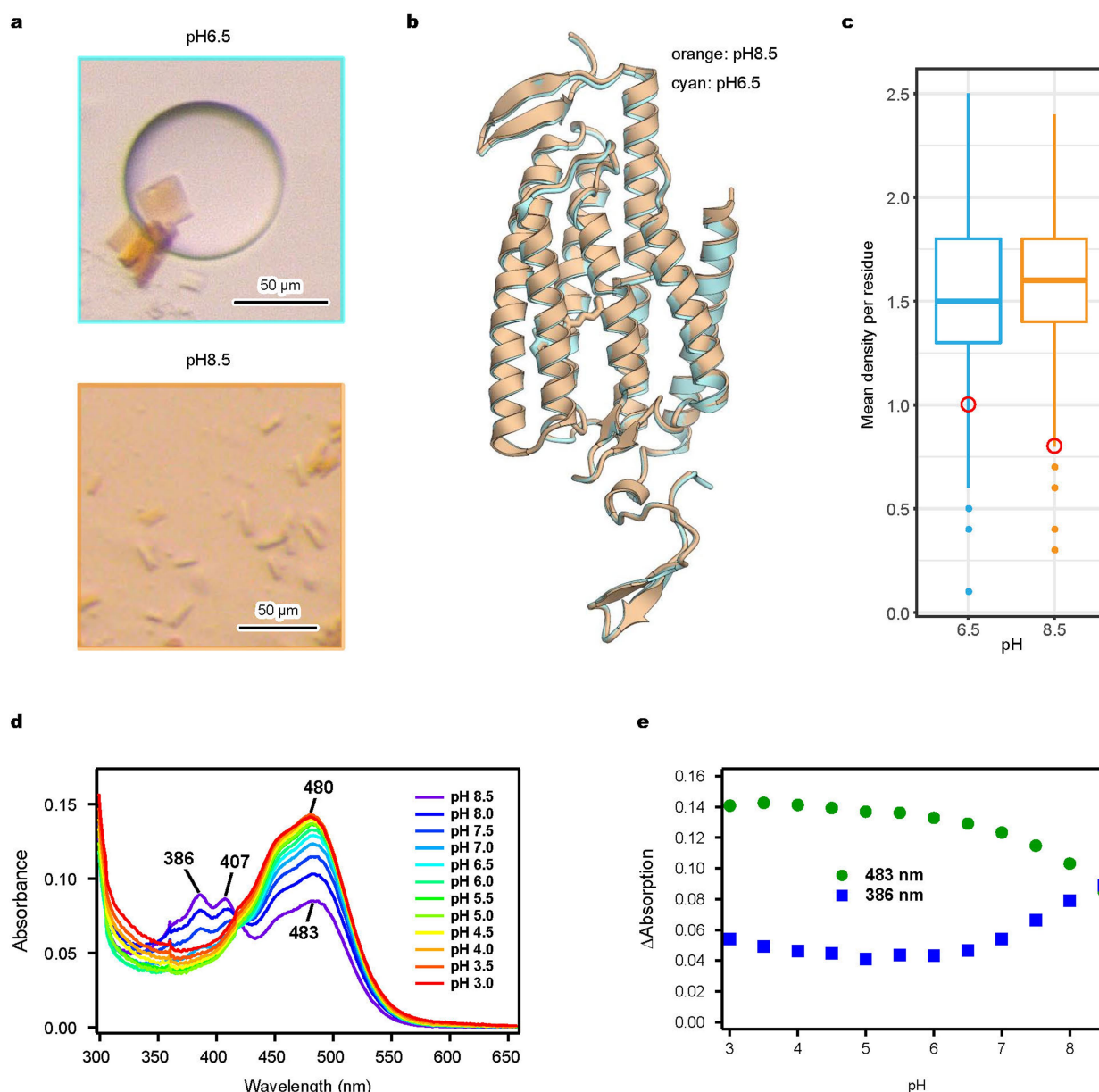
Data availability. The protein coordinates and atomic structure factors of iC++ at pH 8.5 and pH 6.5 have been deposited in the Protein Data Bank (PDB) under accession number 6CSN (pH 8.5) and 6CSO (pH 6.5), respectively. The raw diffraction images have been deposited in the SGrid data repository (ID: 570 for pH 8.5 and 571 for pH 6.5). All other data are available from the corresponding authors upon reasonable request.

22. Yamashita, K., Hirata, K. & Yamamoto, M. KAMO: towards automated data processing for microcrystals. *Acta Crystallogr. D Struct. Biol.* **74**, 441–449 (2018).
23. Kabsch, W. Xds. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
24. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658–674 (2007).
25. Murshudov, G. N. et al. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).
26. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
27. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
28. Gradinaru, V. et al. Molecular and cellular approaches for diversifying and extending optogenetics. *Cell* **141**, 154–165 (2010).
29. Zhang, L. & Hermans, J. Hydrophilicity of cavities in proteins. *Proteins* **24**, 433–438 (1996).
30. Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. OPM: orientations of proteins in membranes database. *Bioinformatics* **22**, 623–625 (2006).
31. Betz, R. *Dabble* v.2.6.3 <https://doi.org/10.5281/zenodo.836914> (2004).
32. Best, R. B., Mittal, J., Feig, M. & MacKerell, A. D. Jr. Inclusion of many-body effects in the additive CHARMM protein CMAP potential results in enhanced cooperativity of alpha-helix and beta-hairpin formation. *Biophys. J.* **103**, 1045–1051 (2012).
33. Best, R. B. et al. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain χ_1 and χ_2 dihedral angles. *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).
34. Huang, J. & MacKerell, A. D., Jr. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145 (2013).
35. Klauda, J. B. et al. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* **114**, 7830–7843 (2010).
36. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
37. Zhu, S., Brown, M. F. & Feller, S. E. Retinal conformation governs pK_a of protonated Schiff base in rhodopsin activation. *J. Am. Chem. Soc.* **135**, 9391–9398 (2013).
38. Salomon-Ferrer, R., Gotz, A. W., Poole, D., Le Grand, S. & Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.* **9**, 3878–3888 (2013).
39. Case, D. A. et al. *AMBER 2017* (University of California, San Francisco, 2017).
40. Hopkins, C. W., Le Grand, S., Walker, R. C. & Roitberg, A. E. Long-time-step molecular dynamics through hydrogen mass repartitioning. *J. Chem. Theory Comput.* **11**, 1864–1874 (2015).
41. Roe, D. R. & Cheatham, T. E. III. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
42. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
43. Vanommeslaeghe, K. et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671–690 (2010).
44. Steinbrecher, T., Latzer, J. & Case, D. A. Revised AMBER parameters for bioorganic phosphates. *J. Chem. Theory Comput.* **8**, 4405–4412 (2012).
45. Paxinos, G., Franklin, K. B. J. *The Mouse Brain in Stereotaxic Coordinates* 2nd edn (Academic Press, San Diego, 2001).
46. Zhang, F. et al. Multimodal fast optical interrogation of neural circuitry. *Nature* **446**, 633–639 (2007).
47. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
48. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32**, W665–7 (2004).



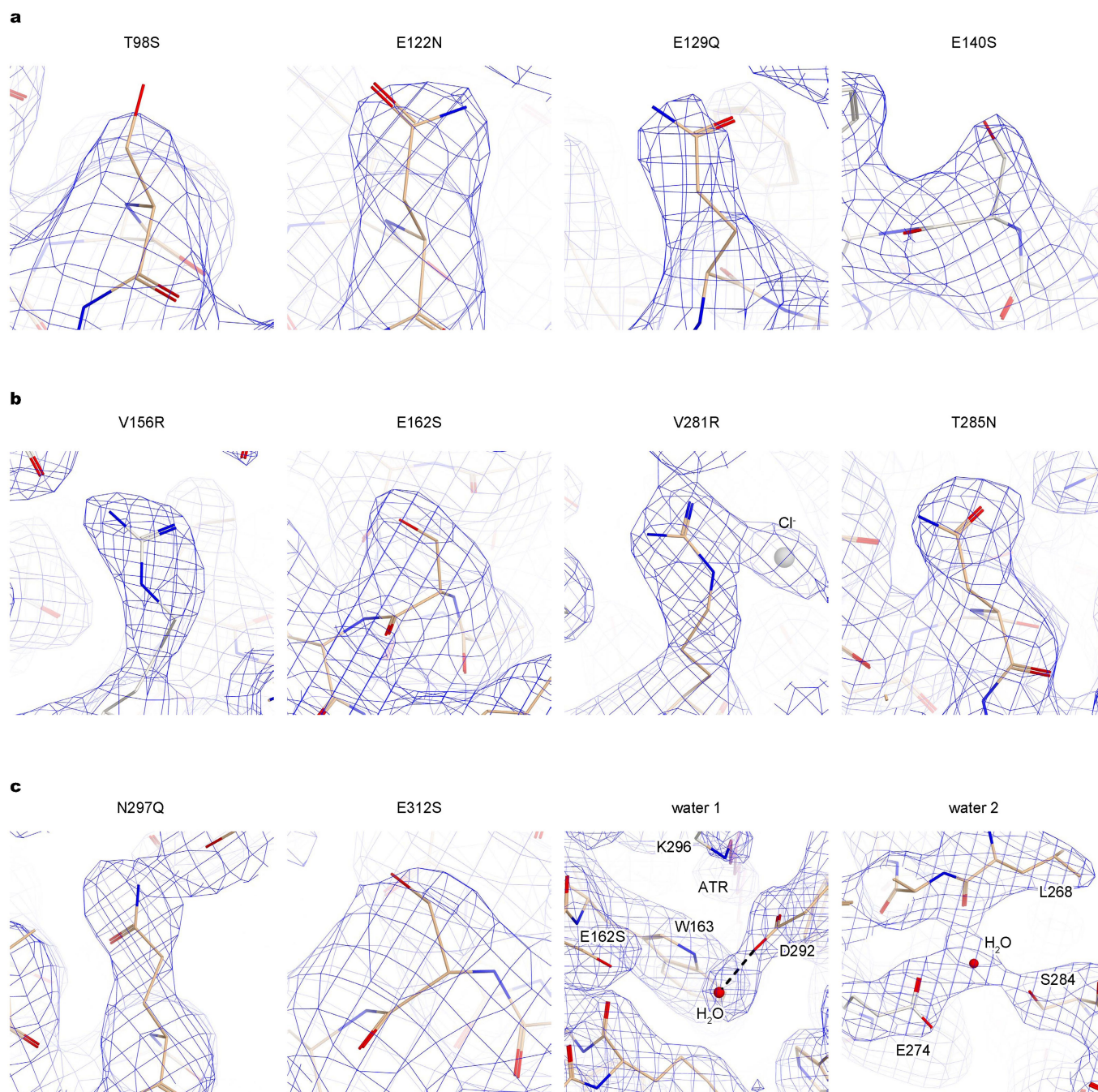
Extended Data Fig. 1 | Spectroscopic and structural characterization of iC++. **a**, Sequence alignment. iC++ mutations and corresponding residues in the other opsins are highlighted. **b**, **c**, Transient absorption spectra of iC++ (**b**) and time traces of the absorption changes (**c**) at specific probe wavelengths. **d**, Photocycle schematic for iC++, determined from analysis of results shown in **b** and **c**. **e**, Overlay of the M-intermediate state measured in **b** (blue) and a flash-induced

photocurrent generated by iC++ in a HEK293 cell (red) by flashing 100 μs , 470 nm light at 2 mW mm^{-2} . The electrophysiology experiment was repeated independently 4 times with similar results, and the spectroscopy experiment was performed once. **f**, Table describing data collection and refinement statistics of iC++ in pH 6.5 and pH 8.5 conditions. Datasets were collected from 16 (pH 8.5) and 27 (pH 6.5) crystals. Values in parentheses are for the highest resolution shell.



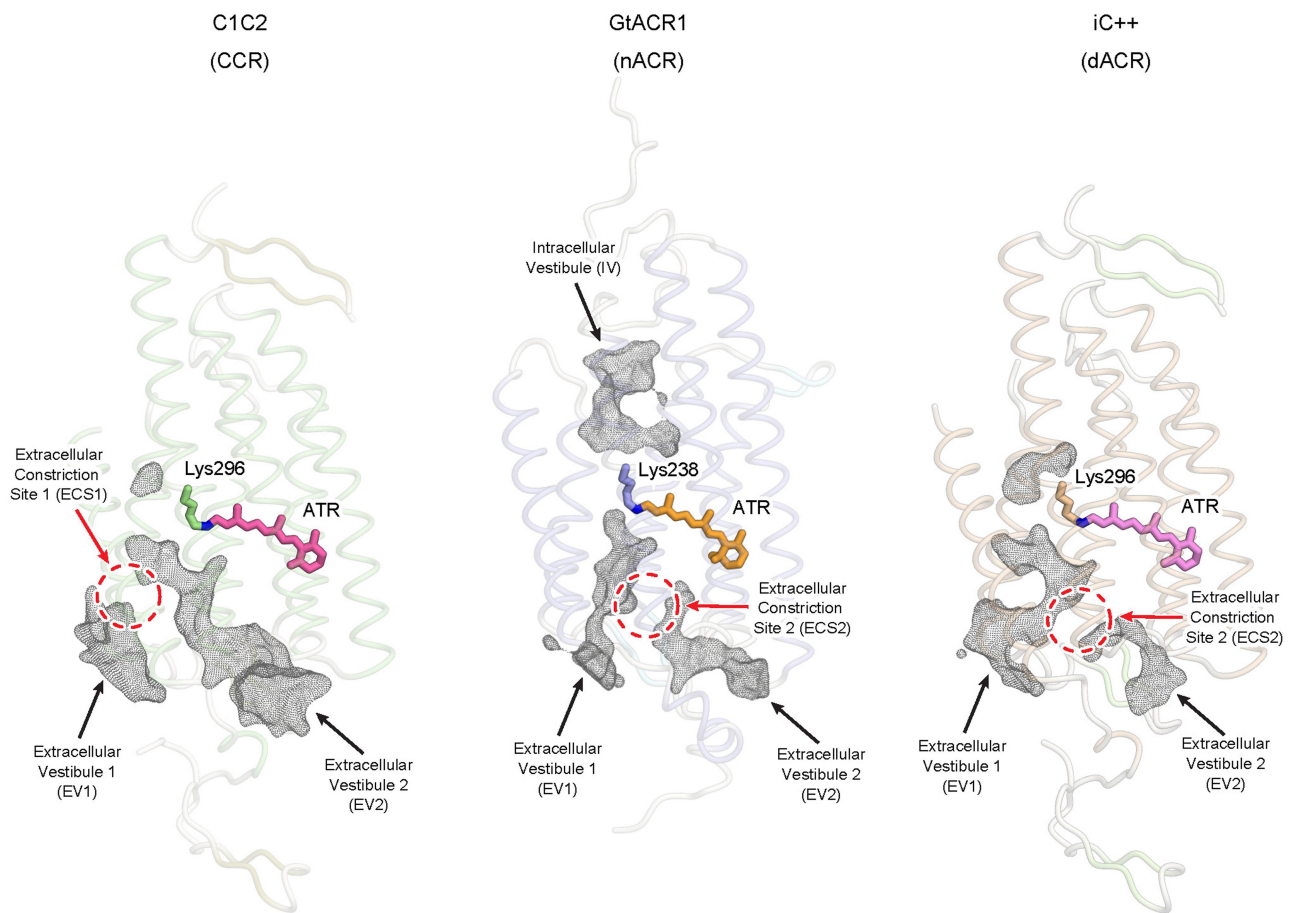
Extended Data Fig. 2 | Structural basis of pH dependence. **a**, Bright-field images of iC++ crystals formed under pH 6.5 (top) and pH 8.5 (bottom) conditions. **b**, Overlay of iC++ crystal structures at pH 6.5 (cyan) and pH 8.5 (orange). **c**, Comparison of calculated mean electron density of ATR atoms (red circle) to the distribution corresponding to protein residues (box-and-whisker plots) at pH = 6.5 (cyan) or 8.5 (orange). Note lower mean electron density of ATR at pH 8.5 despite higher electron density of the overall structure. Mean values at atomic positions of the $2F_o - F_c$ map were calculated using `phenix.get_cc_mtz_pdb` in Phenix²⁶. Box plots show median (centre), first and third interquartile ranges, minimum and

maximum. Sample size (number of residues used in calculation), $n = 282$ for pH 8.5, and 283 for pH 6.5. Electron-density values were normalized in the model region, and maps were generated at 3.2 Å resolution. **d**, **e**, Absorption spectrum of iC++ measured over the range from pH 3.0 to pH 8.5 (**d**), and 483 nm and 386 nm absorbance traces collected across the measured pH range (**e**). Absorbance was measured for every 0.5 pH unit change while the sample was titrated by HCl. Note increased absorbance at 386 nm and decreased absorbance at 483 nm under alkaline conditions.



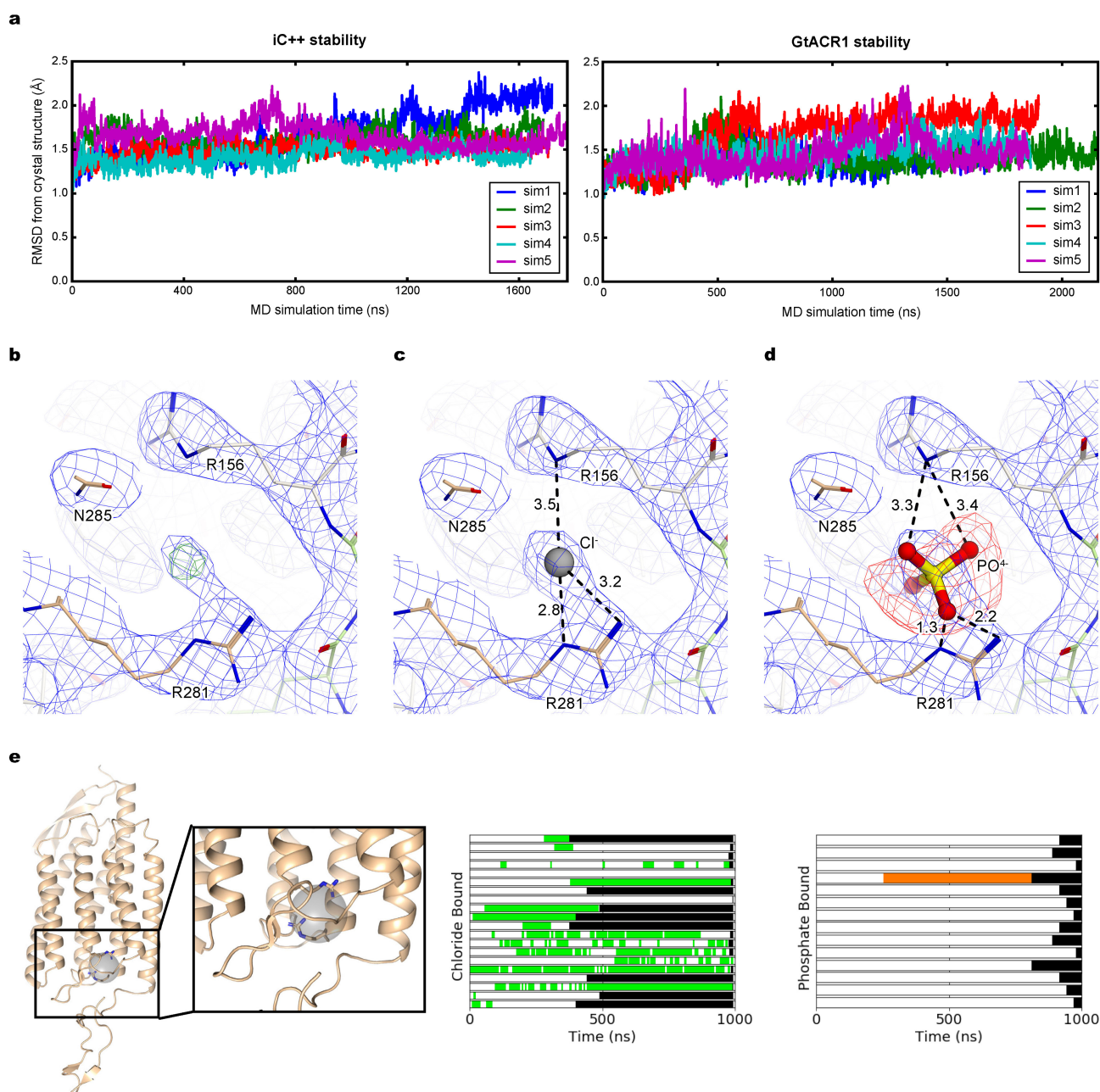
Extended Data Fig. 3 | Electron densities of side chains and putative water molecules of iC++ structure at pH 8.5. Shown are $2F_o - F_c$ maps (blue mesh, contoured at 1σ) for ten introduced mutations, and for two

putative water molecules as labelled. All ten mutations have well-resolved electron densities to reliably position structural features, which support the interpretation of the roles of introduced mutations.



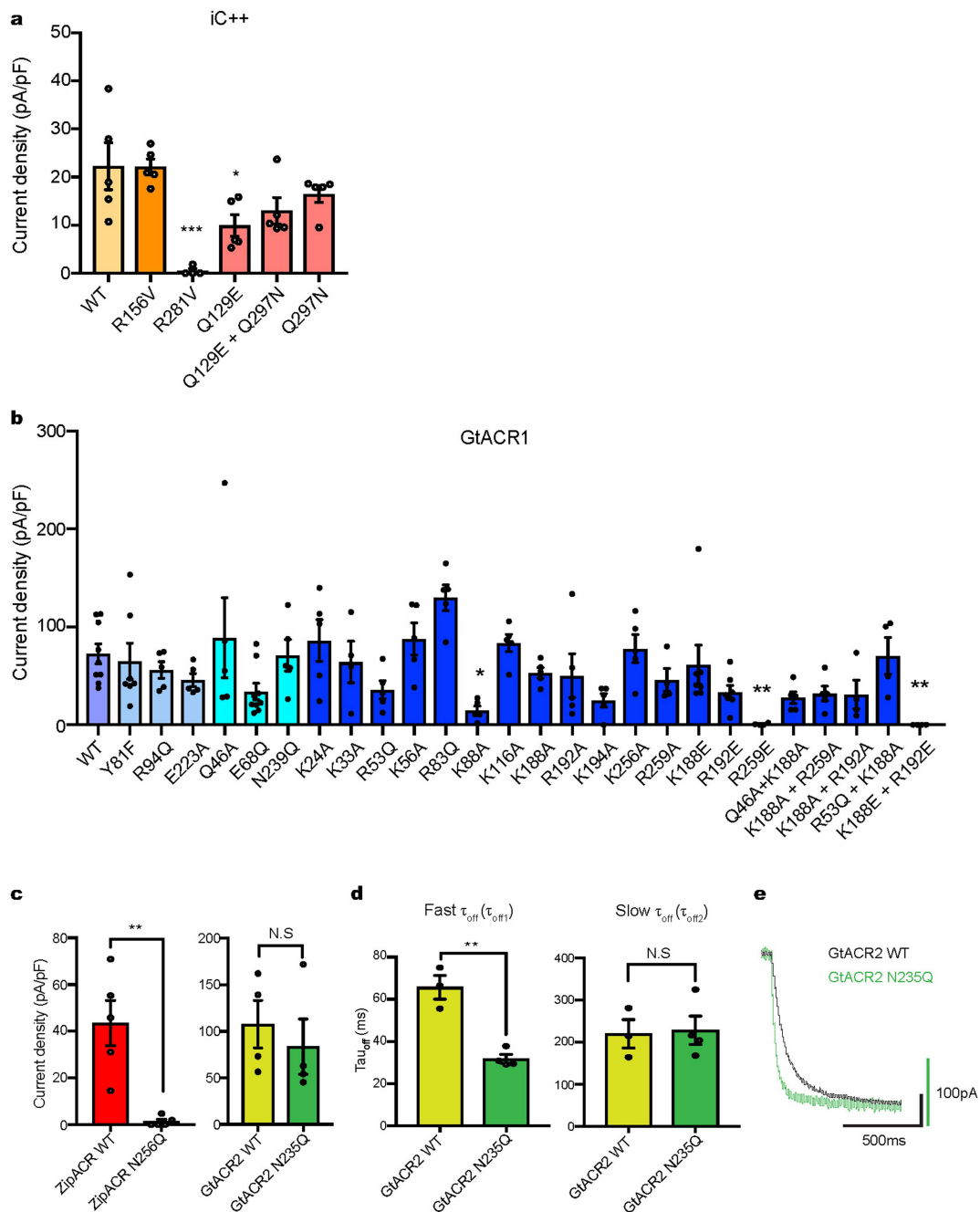
Extended Data Fig. 4 | Ion-conducting pathways. Ion-conducting pathways of C1C2 (left), GtACR1 (middle) and iC++ (right). Black meshes represent the extracellular and intracellular vestibules. Red dashed

circles represent ECSs. Note that EV1 and EV2 of iC++ are structurally more similar to the extracellular vestibules of GtACR1 than those of C1C2.



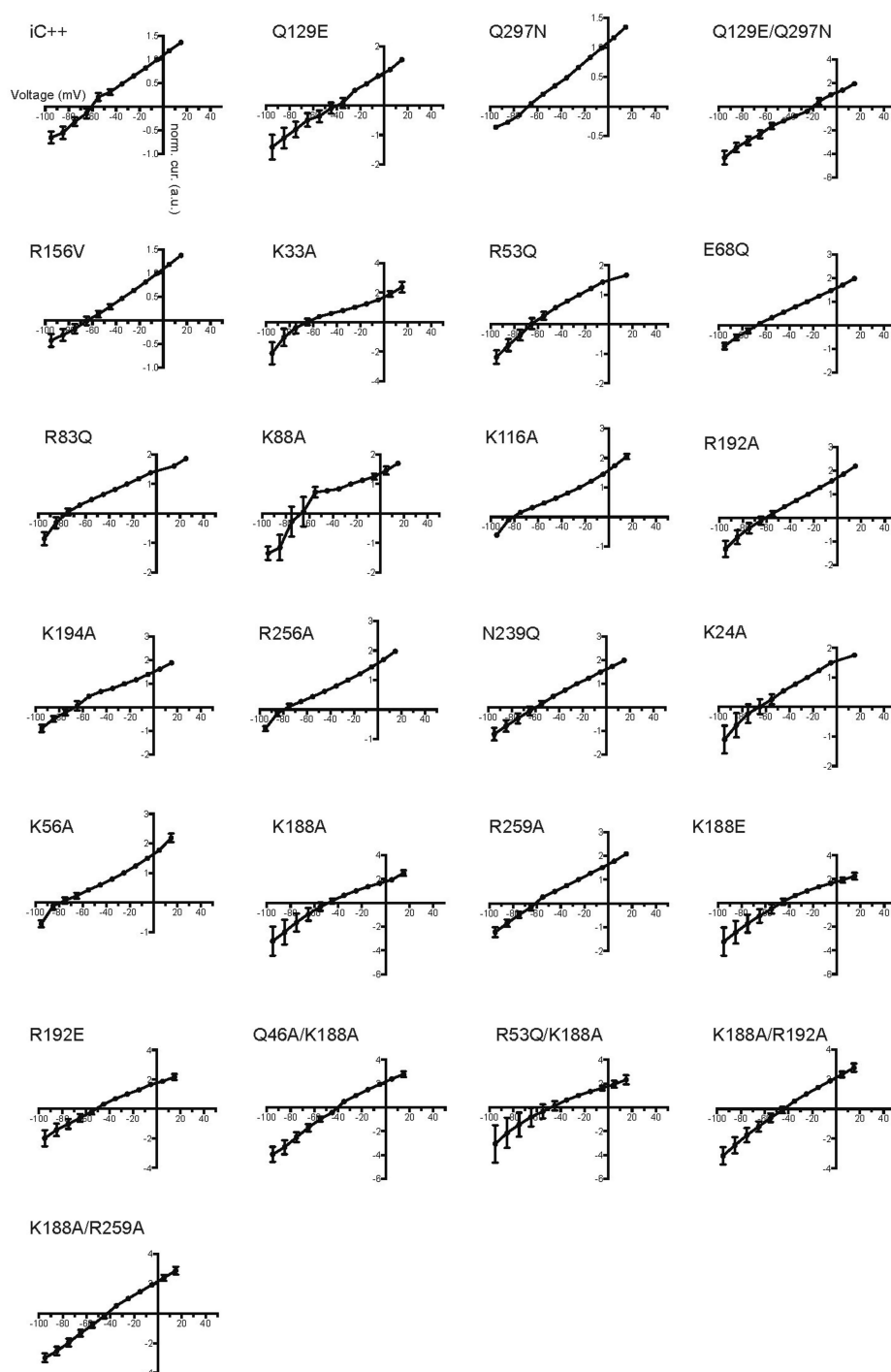
Extended Data Fig. 5 | Computational and crystallographic estimation of the potential chloride-binding site in iC++ at pH 8.5. **a**, The r.m.s.d. of the atomic positions of α -carbons between the crystallographic structures and structures from molecular dynamics simulations was used to assess the stability of the crystal structures in a membrane environment and the quality of our molecular dynamics simulation parameters. The r.m.s.d. values computed from five independent simulations, each over 1500 ns in length, of iC++ (left) and GtACR1 (right), average less than 2 Å, supporting the stability of the crystal structures. In our calculations, we removed the flexible C-terminal tails of both proteins (residue 312 and 251 onwards in iC++ and GtACR1, respectively) and the first eight residues of GtACR1. **b**, $2F_o - F_c$ map (blue mesh, contoured at 1σ) and $F_o - F_c$ omit map (green mesh, contoured at 4σ) for electron density around the putative chloride-binding site. **c**, $2F_o - F_c$ map for the same

site, with a chloride ion (**c**) and with a phosphate (**d**). Note the strong negative peak observed around phosphate ion. Numbers indicate distance between two atoms connected by dashed lines. **e**, Chlorides frequently occupy a site near Arg281 and Arg156 on the extracellular side of iC++, during molecular dynamics simulation. The left panel depicts this binding site as a grey sphere from the perspective looking out from the dimer interface with the extracellular side at the top. Arg281 and Arg156 are shown as sticks. The other two panels show intervals of simulation in which a chloride (middle) or phosphate (right) are present in the binding site. Each horizontal bar represents an individual replicate and monomer. Time periods where a chloride or phosphate bound are coloured green and orange, respectively, and the period after the end of the simulation is coloured black.



Extended Data Fig. 6 | Photocurrent densities and kinetics of wild-type and mutant ACRs. a, b, Summary of current densities of wild-type (WT) *iC++* and its mutants (a) and wild-type *GtACR1* and its mutants (b) from this study. Data are mean \pm s.e.m.; one-way ANOVA followed by Dunnett's test. $*P < 0.05$, $**P < 0.01$ and $***P = 0.0001$. c, Photocurrent comparisons of wild-type ZipACR and N235Q mutant (left) and wild-type *GtACR2* and N235Q mutant (right). Data are means \pm s.e.m.; two-tailed *t*-test, $**P = 0.0025$. d, Fast (left) and slow (right) off-kinetics comparison

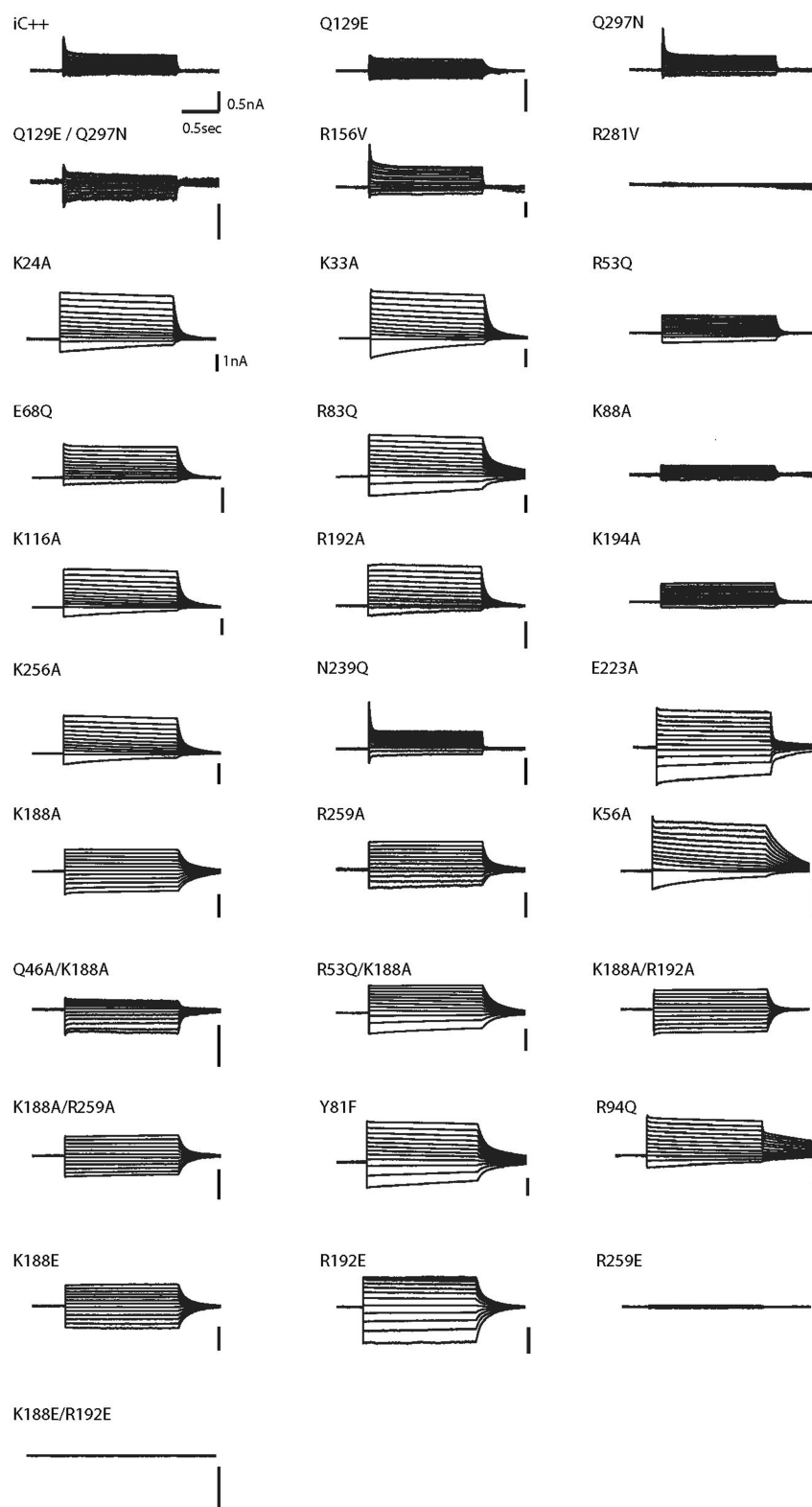
of wild-type *GtACR2* and the N235Q mutant, Data are mean \pm s.e.m.; two-tailed *t*-test, $**P = 0.0014$. e, Example traces of wild-type *GtACR2* and the N235Q mutant expressed in HEK293 cells by lipofectamine transfection, measured at -10 mV holding potential in voltage-clamp. Traces were recorded while cells were stimulated with 1.5 s of 1.0 mW mm $^{-2}$ irradiance at 470 nm for *iC++* and *GtACR2* and 513 nm for *GtACR1* and ZipACR. Sample size (number of cells) for each experiment is indicated next to label in parentheses.



Extended Data Fig. 7 | Current–voltage relationships of ACR mutants.

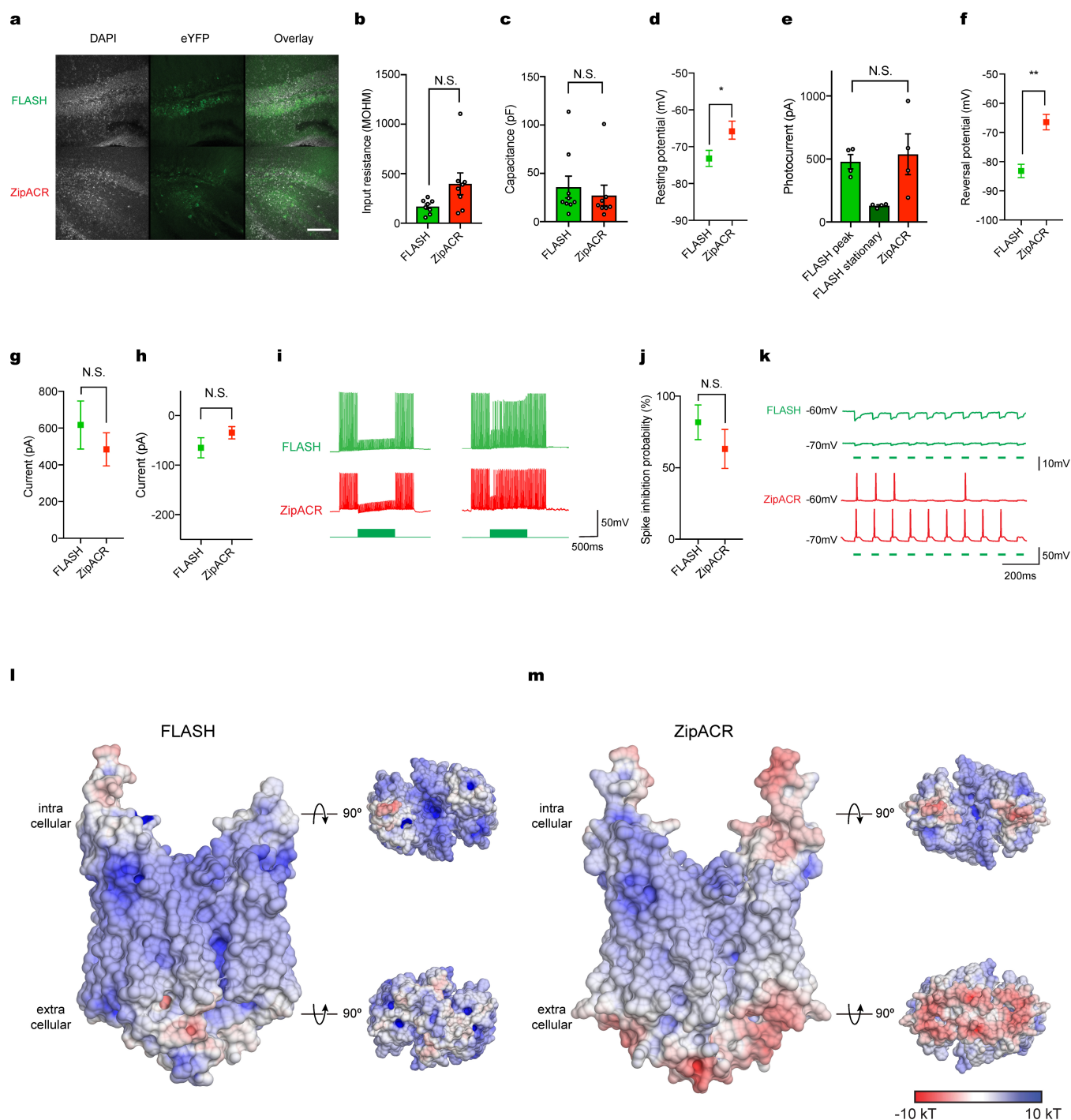
Current–voltage relationships from -95 mV to $+15$ mV were determined from light-evoked current amplitude at the indicated holding potentials. Each measurement was normalized to the current amplitude measured at -5 mV for iC++ variants and -25 mV for *GtACR1* variants. Values are means and s.e.m. For iC++ variants: $n = 8$ for wild type, and $n = 5$ for the other iC++ variants. For *GtACR1* variants: $n = 9$ for E68Q, $n = 7$ for

R192E, $n = 5$ for K24A, K33A, K33E, R53Q, K56A, R83Q, K116A, K188A, K188E, R192A, K256A, Q46A/K188A, K188A/R259A, and $n = 4$ for the other variants. HEK293 cells expressing proteins through lipofectamine transfection method were recorded while stimulated by 1.5 s of 1.0 mW/mm² irradiance at 470 nm for iC++ and 513 nm for *GtACR1*. The first five graphs are from the iC++ backbone, and the rest are *GtACR1* mutations.



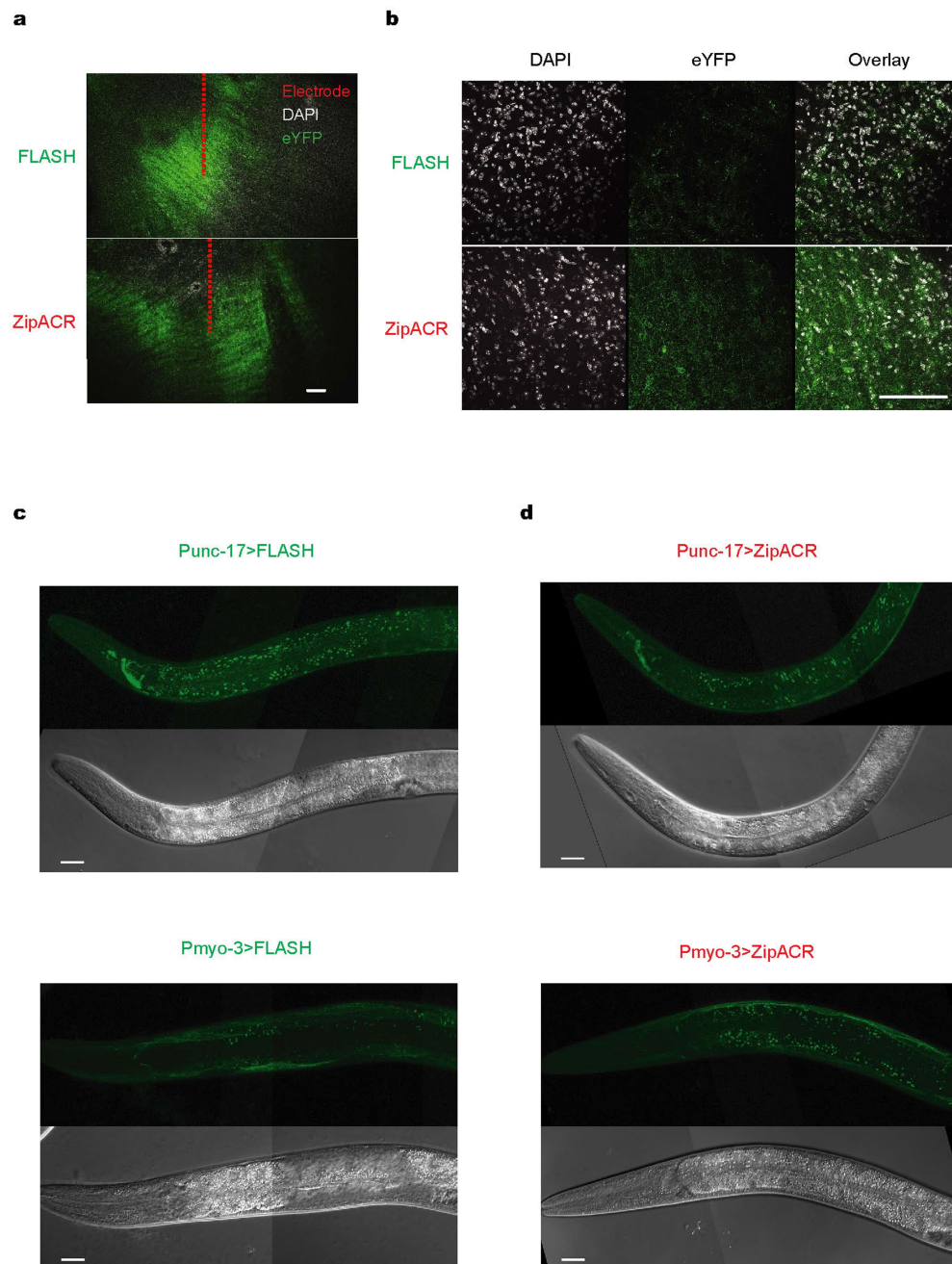
Extended Data Fig. 8 | Representative traces for current-voltage measurements. Voltage-clamp traces corresponding to current-voltage relationships shown in Extended Data Fig. 7, collected from -95 mV to $+15$ mV in steps of 10 mV. HEK293 cells, transfected using the

lipofectamine transfection method, were recorded while stimulated by 1.5 s of 1.0 mW mm $^{-2}$ irradiance at 470 nm for iC++ and 513 nm for GtACR1. The first six traces are from the iC++ backbone, and the rest are GtACR1 mutations.



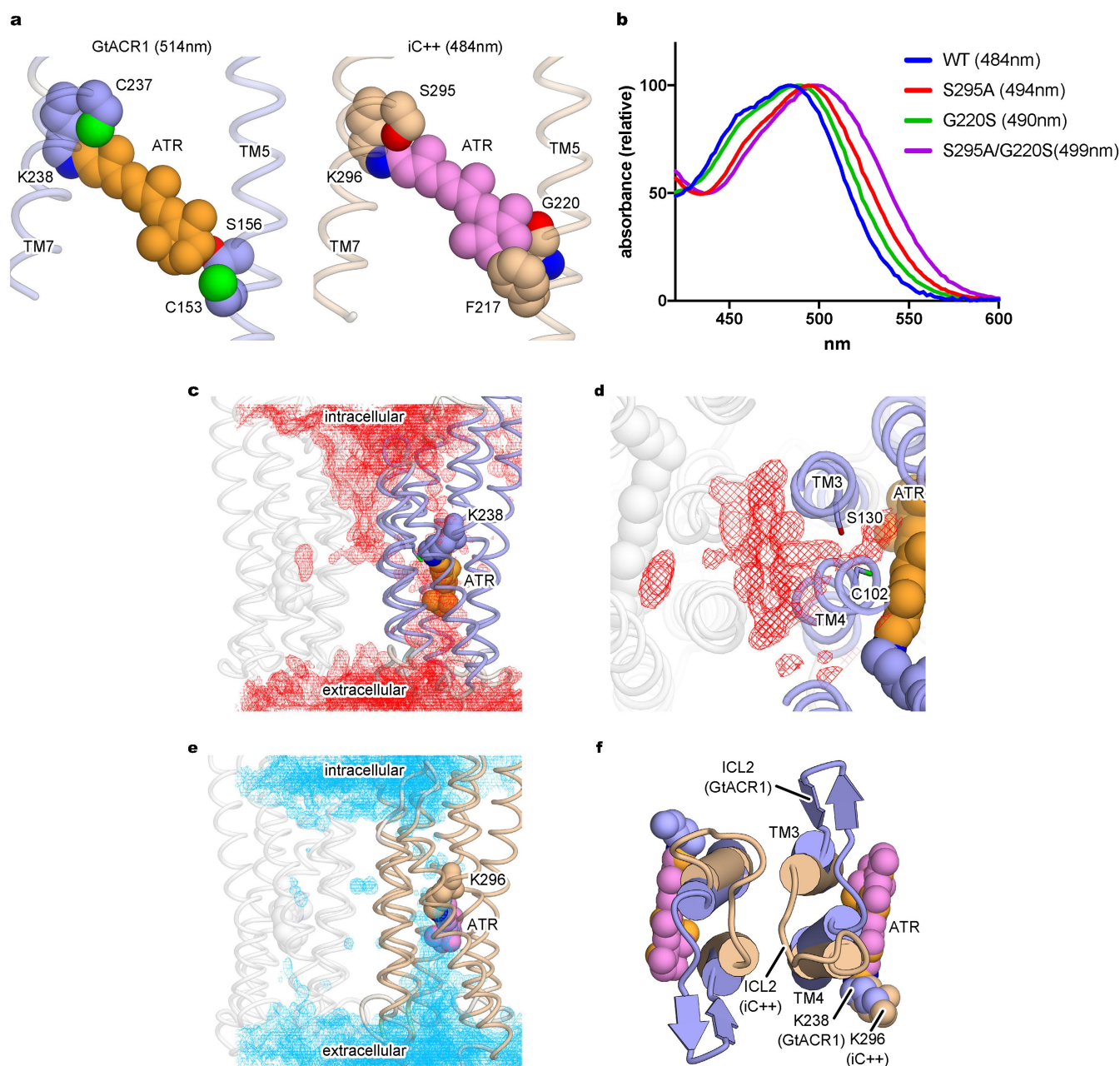
Extended Data Fig. 9 | Electrophysiology and structural comparison of FLASH and ZipACR. **a**, Confocal images of FLASH and ZipACR expression in mouse hippocampus 4 weeks after virus injection. Up to 15 slices per animals were collected, all with similar expression level of opsins. Scale bar, 100 μ m. **b–h**, Comparisons of input resistance (**b**), capacitance (**c**), resting potentials (**d**), photocurrent measured at -10 mV (**e**), reversal potentials (**f**), currents injected to elicit action potentials (**g**), and holding currents needed to keep cells at -70 mV (**h**). The steady-state/peak photocurrent ratio of ZipACR was close to 100%, so the stationary current data was not included in **e**. All electrophysiology data were measured from FLASH- and ZipACR-expressing hippocampal neurons in acute slice preparation from animals 4 weeks after virus injection. Data are mean \pm s.e.m.; two-tailed t -test. * $P = 0.044$, ** $P = 0.003$. **i**, Example

traces of neurons expressing FLASH and ZipACR succeeding (left) and failing (right) to inhibit multiple spikes under the pulsed illumination condition. **j**, Summary of FLASH and ZipACR multiple-spike inhibition, corresponding to the experimental paradigm in **i**. Data are mean \pm s.e.m.; two-tailed t -test. $P = 0.32$. **k**, Light-driven spiking was observed from ZipACR expressing neurons (3/8), but not from FLASH expressing ones (0/8). For all electrophysiology experiments, sample size (number of cells) for each experiment is indicated next to label in parentheses. **l**, **m**, Surface electrostatic potentials of homology models of FLASH (**l**) and ZipACR (**m**). Models were built using SWISS-MODEL⁴⁷. The surface is coloured on the basis of electrostatic potential contoured from -10 kT (red) to $+10$ kT (blue). White denotes 0 kT. Surface potential was calculated using PDB2PQR⁴⁸ for both *GtACR1* and *C1C2* models.



Extended Data Fig. 10 | Confocal images of opsin expression in brain slices and worms. a, b, Confocal images showing opsin expression with electrode (a) and in thalamus (b). Scale bar, 100 μm . **c, d,** Confocal images

showing FLASH (c) and ZipACR (d) expression in worms, mediated by *punc-17*- (cholinergic neurons, top) and *pmyo-3*- (muscle, bottom) promoters. Scale bar, 20 μm .



Extended Data Fig. 11 | Differences in polarity and hydration of retinal binding pockets in *iC++* and *GtACR1*. **a**, Structures of the retinal binding pockets of *GtACR1* (left) and *iC++* (right) in van der Waals representations. **b**, Absorption spectra of wild-type *iC++*, S295A, G220S and S295A/G220S mutants. Water density map at the dimer interface of *GtACR1*, viewed parallel to and from within the membrane (**c**) and viewed from the intracellular side (**d**) during molecular dynamics simulation,

contoured at probability density of 0.016 molecules per \AA^3 . **e**, Corresponding water density map at the dimer interface for *iC++* viewed parallel to and from within the membrane, contoured at probability density of 0.016 molecules per \AA^3 . **f**, Superimposed structures of *iC++* (beige) and *GtACR1* (blue) at the dimer interfaces, viewed from the intracellular side. ATR molecules are shown as sphere models.

Superluminal motion of a relativistic jet in the neutron–star merger GW170817

K. P. Mooley^{1,2,10*}, A. T. Deller^{3,4,10}, O. Gottlieb^{5,10*}, E. Nakar⁵, G. Hallinan², S. Bourke⁶, D. A. Frail¹, A. Horesh⁷, A. Corsi⁸ & K. Hotokezaka⁹

The binary neutron-star merger GW170817¹ was accompanied by radiation across the electromagnetic spectrum² and localized² to the galaxy NGC 4993 at a distance³ of about 41 megaparsecs from Earth. The radio and X-ray afterglows of GW170817 exhibited delayed onset^{4–7}, a gradual increase⁸ in the emission with time (proportional to $t^{0.8}$) to a peak about 150 days after the merger event⁹, followed by a relatively rapid decline^{9,10}. So far, various models have been proposed to explain the afterglow emission, including a choked-jet cocoon^{4,8,11–13} and a successful-jet cocoon^{4,8,11–18} (also called a structured jet). However, the observational data have remained inconclusive^{10,15,19,20} as to whether GW170817 launched a successful relativistic jet. Here we report radio observations using very long-baseline interferometry. We find that the compact radio source associated with GW170817 exhibits superluminal apparent motion between 75 days and 230 days after the merger event. This measurement breaks the degeneracy between the choked- and successful-jet cocoon models and indicates that, although the early-time radio emission was powered by a wide-angle outflow⁸ (a cocoon), the late-time emission was most probably dominated by an energetic and narrowly collimated jet (with an opening angle of less than five degrees) and observed from a viewing angle of about 20 degrees. The imaging of a collimated relativistic outflow emerging from GW170817 adds substantial weight to the evidence linking binary neutron-star mergers and short γ -ray bursts.

Our very long-baseline interferometry (VLBI) observations with the High Sensitivity Array (HSA)—which consists of the Very Long Baseline Array (VLBA), the Karl G. Jansky Very Large Array (VLA) and the Robert C. Byrd Green Bank Telescope (GBT)—75 and 230 days after the GW170817 merger event (mean epochs; see Methods) indicate that the centroid position of the radio counterpart of GW170817 changed from a right ascension of RA = 13 h 09 min 48.068638(8) s and declination of dec. = $-23^{\circ} 22' 53.3909(4)''$ to RA = 13 h 09 m 48.068831(11) s and dec. = $-23^{\circ} 22' 53.3907(4)''$ between these epochs (1σ uncertainties in the last digits are given in parentheses). This implies a positional offset between the two observations of $2.67 \pm 0.19 \pm 0.21$ mas in RA and $0.2 \pm 0.6 \pm 0.7$ mas in dec. (1σ uncertainties; statistical and systematic, respectively; see Methods). This corresponds to a mean apparent velocity of the source of the radio counterpart along the plane of the sky of $\beta_{\text{app}} = 4.1 \pm 0.5$, where β_{app} is in units of the speed of light, c (1σ , including the uncertainty in the source distance). Offset positions of the radio source and the positional uncertainties at both VLBI epochs are shown in Fig. 1. Our VLBI data are consistent with the source being unresolved at both day 75 and day 230. Given the VLBI angular resolution and the signal-to-noise ratio of the detection, this puts an upper limit on the size of the source in both epochs of about 1 mas (0.2 pc at the distance of NGC 4993) in the direction parallel to its motion and 10 mas perpendicular to its motion (see Methods).

The substantial proper motion of the radio source immediately rules out isotropic ejecta models^{21–23} for the radio (and X-ray) afterglow, which predict proper motion close to zero, and argues in favour of highly anisotropic ejecta (consistent with jet models). If the ejecta are bipolar, then one of the components is relativistically beamed into our line of sight.

Although superluminal motion is seen frequently in active galactic nuclei and micro-quasars, it is extremely rare in extragalactic explosive transients. Superluminal motion has been measured in only one such transient: the long-duration γ -ray burst GRB 030329²⁴. GRB 030329 had a measured superluminal expansion of $\beta_{\text{app}} \approx 3$ –5, but no proper motion, whereas GW170817 has measured proper motion, but no expansion. Although both were relativistic events of comparable energies, these differences suggest different geometries and/or viewing angles.

The apparent velocity and size of a source moving at relativistic speeds, such as the radio counterpart of GW170817, differs from its actual velocity and size. The image of a point source, moving at a Lorentz factor Γ and viewed at an angle θ , is point-like and has a maximal apparent velocity of $\beta_{\text{app}} = \Gamma$, which is obtained when $\theta = 1/\Gamma$. On the other hand, the maximal centroid velocity of an extended source with a uniform Γ is less than Γ , and its image size increases²⁵ with the source size and with Γ . An extreme example of the latter case is a spherically symmetric source expanding isotropically. In such a case, the image is a ring with a radius that increases at a velocity Γ with no centroid motion. The centroid velocity may also be affected in cases where we see different regions of the outflow at different times²⁶ (that is, a pattern motion).

Using this information, we examine the results from the VLBI data and the radio light curve to derive analytical constraints on the geometry and size of the radio source. We assume that the ejecta are axis-symmetric, so that θ_{obs} is the viewing angle and θ_s is the average angular size of the source that dominates the emission between 75 and 230 days after the merger (both with respect to the symmetry axis). If the source is compact ($\theta_s \lesssim \theta_{\text{obs}} - \theta_s$), its size and possible pattern motion have only a small effect on the observed radiation and so we can use the point-source approximation. In all of the highly aspherical models suggested, the energy density increases towards the axis of symmetry, implying that during the peak of the light curve the emission is dominated by a region at $(\theta_{\text{obs}} - \theta_s) \approx 1/\Gamma$. Using the point-source approximation, this implies that between the two observations the source is observed at an angle of $(\theta_{\text{obs}} - \theta_s) \approx 1/\beta_{\text{app}} \approx 0.25$ rad and its Lorentz factor is $\Gamma \approx \beta_{\text{app}} \approx 4$. If the source is extended ($\theta_s \gg \theta_{\text{obs}} - \theta_s$), then to achieve the observed apparent velocity the source should have $\Gamma > 4$ and possibly $\theta_{\text{obs}} - \theta_s < 0.25$ rad.

There are several strong lines of evidence that suggest that the source is compact. First, it is very compact in our VLBI observations, and is consistent with being unresolved. Second, the observed flux depends

¹National Radio Astronomy Observatory, Socorro, NM, USA. ²Caltech, Pasadena, CA, USA. ³Centre for Astrophysics and Supercomputing, Swinburne University of Technology, Hawthorn, Victoria, Australia. ⁴ARC Centre of Excellence for Gravitational Wave Discovery (OzGrav), Hawthorn, Victoria, Australia. ⁵The Raymond and Beverly Sackler School of Physics and Astronomy, Tel Aviv University, Tel Aviv, Israel. ⁶Department of Space, Earth and Environment, Chalmers University of Technology, Onsala Space Observatory, Onsala, Sweden. ⁷Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem, Israel. ⁸Department of Physics and Astronomy, Texas Tech University, Lubbock, TX, USA. ⁹Department of Astrophysical Sciences, Princeton University, Princeton, NJ, USA. ¹⁰These authors contributed equally: K. P. Mooley, A. T. Deller, O. Gottlieb. *e-mail: kunal@astro.caltech.edu; oregottlieb@gmail.com

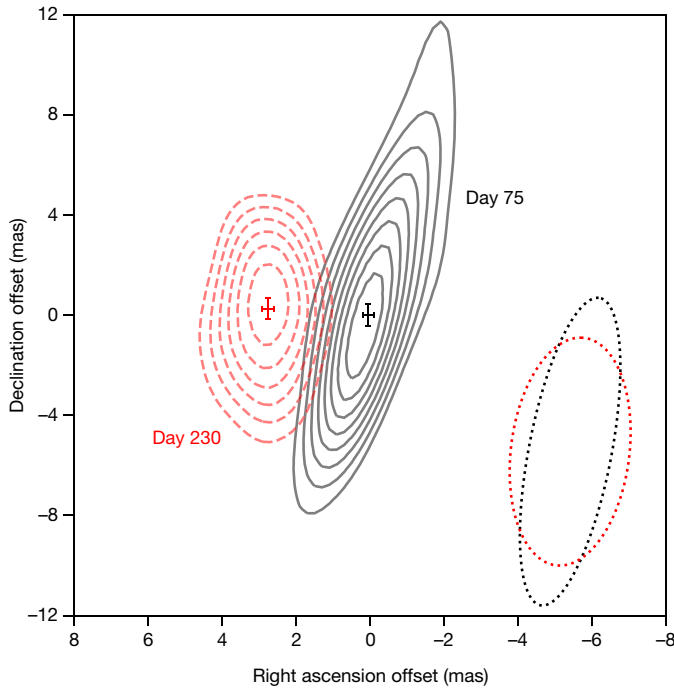


Fig. 1 | Proper motion of the radio counterpart of GW170817. The offset positions of the centroid (shown by 1σ error bars) and 3σ – 12σ contours of the radio source detected 75 days (black) and 230 days (red) after the merger event using VLBI at 4.5 GHz. The two VLBI epochs have image root-mean-square noise of $5.0 \mu\text{Jy beam}^{-1}$ and $5.6 \mu\text{Jy beam}^{-1}$ (natural weighting), respectively, and the peak flux densities of GW170817 are $58 \mu\text{Jy beam}^{-1}$ and $48 \mu\text{Jy beam}^{-1}$, respectively. The radio source is consistent with being unresolved at both epochs. The shapes of the synthesized beams for the images from each epoch are shown as dotted ellipses in the lower right corner. The proper-motion vector of the radio source has a magnitude of $2.7 \pm 0.3 \text{ mas}$ and a position angle of $86^\circ \pm 18^\circ$, measured over 155 days.

very strongly on Γ (as roughly $\Gamma^{1.4}$), which implies that on day 150 the Lorentz factor of the radio source was¹⁹ less than about 5. Last, and most constraining, is the rapid turnover around the peak of the radio light curve and the very fast decline that follows $F_\nu \propto t^{-2}$ after day 200, where F_ν is the flux density and t is the time in the observer frame (K.P.M. et al., manuscript in preparation). The shape of the peak and the following decline depends on the ratio $\theta_s/(\theta_{\text{obs}} - \theta_s)$. A smaller ratio results in a narrower peak, and if $\theta_s \gg \theta_{\text{obs}} - \theta_s$ the decay is expected to be¹⁹ at first roughly linear in time, whereas if $\theta_s \ll \theta_{\text{obs}} - \theta_s$ the flux decay after the peak is predicted to behave as roughly $F_\nu \propto t^{-p}$, where the radio spectrum dictates^{8,12,16} that $p \approx 2.16$. We conclude that the combination of the image and the light curve indicate that around the peak, at day 150, the emission is most probably dominated by a narrow component with $\theta_s \ll 0.25 \text{ rad}$ and $\Gamma \approx 4$, which is observed at an angle of $\theta_{\text{obs}} - \theta_s \approx 0.25 \text{ rad}$ (in contrast to the emission during the first month or two, which was most probably dominated by cocoon emission from angles larger than θ_s).

The constraints derived above strongly disfavour an uncollimated choked jet, where the jet has a wide opening angle and does not successfully escape the neutron-rich material ejected dynamically during the merger (that is, it is choked and so does not contain a relativistic narrow core). A narrowly collimated choked jet may generate an outflow with a narrow high-energy core, but it is hard to obtain a Lorentz factor that is high enough without a fine tuning of the location where the jet is choked. In contrast to all other models, the successful-jet model predicts a structure that can easily satisfy the constraints of the image and the light curve. In this model, the gradual rise is generated by cocoon emission and the peak is observed when the core of the successful jet decelerates and starts to dominate the emission. The jet opening angle θ_j and its Lorentz factor are those of the source in our

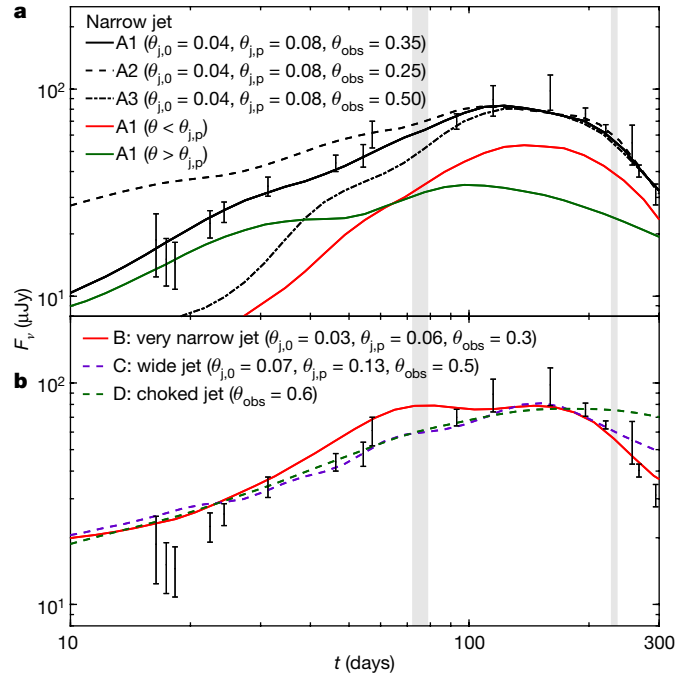


Fig. 2 | Radio, 3-GHz light curves of several representative simulated models. The black error bars (1σ) are the 3-GHz flux density (F_ν) values for GW170817. The grey shaded regions denote the VLBI epochs: 75 and 230 days after the merger. **a**, A narrow jet with an initial opening angle of $\theta_{j,0} = 0.04 \text{ rad}$ (2.3°), total energy of $E = 10^{50} \text{ erg}$ and isotropic equivalent energy of $E_{\text{iso}} = 10^{53} \text{ erg}$ at the core, as observed at three different viewing angles (models A1–A3). For all light curves, we take the energy fraction of accelerated electrons to be $\epsilon_e = 0.1$, assume a power-law index of $p = 2.16$, and vary the energy fraction of the magnetic field ϵ_B and the external density n (which is assumed to be constant in space) to obtain a best fit to the light curve. The opening angle of the jet core at the time of the peak is $\theta_{j,p} = 0.08 \text{ rad}$. The model that gives best fits both for the light curve and the images corresponds to a viewing angle of $\theta_{\text{obs}} = 0.35 \text{ rad}$ ($\epsilon_B = 10^{-4}$, $n = 6 \times 10^{-4} \text{ cm}^{-3}$). The red line shows the contribution of emission from the jet core ($\theta < \theta_{j,p}$) and the green line shows the cocoon emission. The fit to the observations is obtained only in a narrow range of viewing angles. For smaller angles (such as $\theta_{\text{obs}} = 0.25 \text{ rad}$, $\epsilon_B = 2 \times 10^{-4}$, $n = 10^{-4} \text{ cm}^{-3}$) the light curve rises too slowly and the image centroid moves too far, whereas at larger angles (such as $\theta_{\text{obs}} = 0.5 \text{ rad}$, $\epsilon_B = 8 \times 10^{-5}$, $n = 6 \times 10^{-3} \text{ cm}^{-3}$) the light curve rises too quickly and the image centroid motion is too small. **b**, Light curves of three other models. Model B: another narrow jet with a lower energy, $\theta_{j,p} = 0.06 \text{ rad}$, $E = 10^{49} \text{ erg}$ and $E_{\text{iso}} = 2 \times 10^{52} \text{ erg}$ ($\epsilon_B = 4 \times 10^{-5}$, $n = 7 \times 10^{-3} \text{ cm}^{-3}$), at $\theta_{\text{obs}} = 0.3 \text{ rad}$, which provides a reasonable fit to the data. Model C: a wider jet with $\theta_{j,p} = 0.13 \text{ rad}$; even for $\theta_{\text{obs}} = 0.5 \text{ rad}$, the light curve does not decay fast enough to be consistent with the most recent data points, and at this viewing angle the image centroid moves too slowly. Model D: a model of a choked jet; the light curve does not decay fast enough after the peak and the image motion, although superluminal, is very slow compared to the observations. In all of the models that we considered, the spectrum between radio and X-ray frequencies follows a constant power law (cooling and self-absorption do not affect this spectral range) and so models that fit the radio, 3-GHz data fit the entire afterglow observations from radio to X-ray frequencies; see Methods for details.

images around the time of the peak, namely $\theta_j \approx \theta_s$. We can only put a lower limit on the initial Lorentz factor of the jet Γ_0 , because we do not know the deceleration radius (that is, when the transition from the coasting phase to the power-law decline phase took place). All of the observational data can be explained with a narrowly collimated jet with $\Gamma_0 \gtrsim 10$.

To verify the analytical considerations discussed above, and to find tighter constraints on the outflow, we ran a set of relativistic hydrodynamic simulations (see Methods). Our simulations include configurations of choked and successful jets at various opening angles and

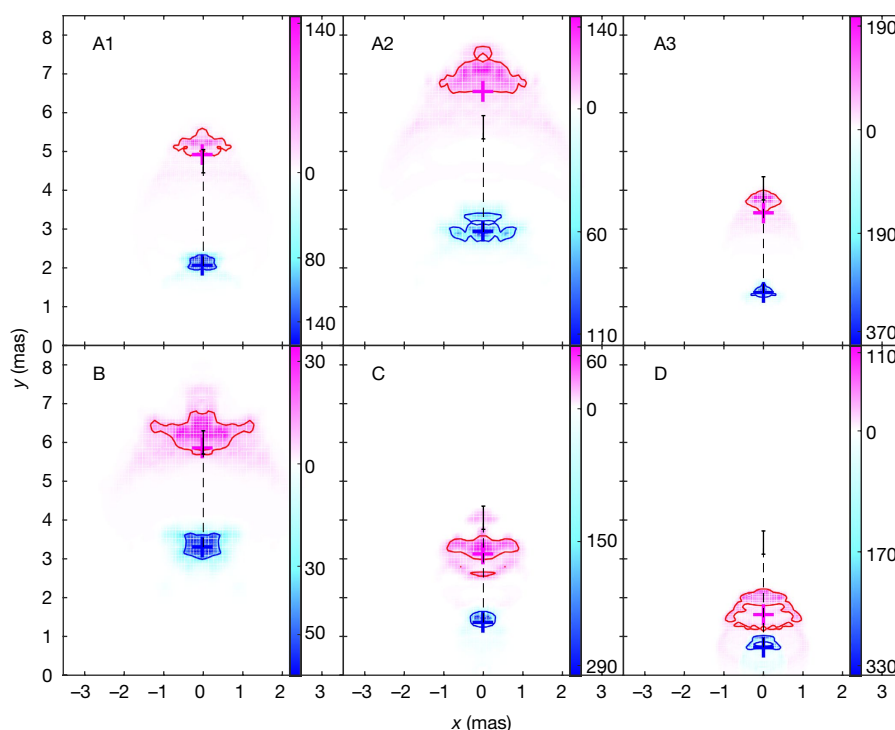


Fig. 3 | Synthetic radio images. Each panel shows two colour maps of the flux density (in units of $\mu\text{Jy mas}^{-2}$), one at day 75 (blue colour palette) and one at day 230 (magenta colour palette), for the models A1–A3, B, C and D shown in Fig. 2. The position at the time of the merger is $x = y = 0$, and the blue and magenta crosses mark the flux centroid at days 75 and 230, respectively. The 50% flux containment contours are also shown at the two epochs. The black dashed lines mark the direction of centroid motion and the black solid segments denote the motion consistent with the VLBI observations within 1σ , 2.7 ± 0.3 mas. Only models A1 and B, which are of narrow jets ($\theta_{\text{jp}} < 0.1$ rad) observed at angles of $\theta_{\text{obs}} = 0.35$ rad and $\theta_{\text{obs}} = 0.3$ rad, respectively, show centroid motions that are consistent with the observations (2.8 mas and 2.6 mas, respectively). These are also the

models that provide the best fits to the light curve. The centroid motion between the two epochs of successful-jet models with larger opening angle ($\theta_{\text{obs}} = 0.5$ rad), A3 and C, is too small (2.1 mas and 1.7 mas, respectively); that of model A2 ($\theta_{\text{obs}} = 0.25$ rad) is too large (3.5 mas). The centroid motion for the choked-jet model, D, is much too small (0.7 mas). In all of the successful-jet models, larger viewing angles lead to more compact images. The observed images were unresolved, with an upper limit on the width parallel to the centroid motion of about 1 mas (1σ). Models A1, A3 and C ($\theta_{\text{obs}} \geq 0.35$ rad) are consistent with this limit, model B ($\theta_{\text{obs}} = 0.3$ rad) is marginal and model A2 ($\theta_{\text{obs}} = 0.25$ rad) is too extended. See Fig. 2 and Methods for further details of the various models and their fitting to the VLBI data.

viewing angles, and include emission from all components of the outflow. In Fig. 2 we show light curves from six different configurations, and in Fig. 3 the corresponding images at days 75 and 230. As expected, we find that in simulations in which the jet is choked, the centroid velocity of the images is too slow to explain the proper motion of GW170817, and the decline of the light curve after the peak is much slower than t^{-2} . Among the successful-jet simulations, those that correspond to jets that moved fast enough; on the other hand, in the simulation of jets that were observed at an angle that is too small ($\theta_{\text{obs}} - \theta_{\text{j}} \lesssim 0.2$ rad), the image centroid moved too fast and/or the source size was too large. The light curve also constrains the geometry, and only simulations with $\theta_{\text{j}}/(\theta_{\text{obs}} - \theta_{\text{j}})$ small enough can fit the rapid transition from a rising light curve to the observed decay. Of all of the configurations that we examined, only extremely narrow jets with $\theta_{\text{j}} < 0.1$ rad that were observed at an angle of $0.2 \text{ rad} < \theta_{\text{obs}} - \theta_{\text{j}} < 0.4$ rad result in emission that is consistent with the light curve and that reproduces the observed motion of the image centroid. Taken together, these results imply that we see a narrow jet with $\theta_{\text{j}} < 0.1$ rad ($< 5^\circ$) from a viewing angle in the range $0.25 \text{ rad} < \theta_{\text{obs}} < 0.50$ rad (14° – 28°). This can be seen, for example, in Figs. 2 and 3, in which the centroid motion for models with viewing angles outside of this range deviate significantly (by more than 2σ ; see Methods) from the observations, and models with wider jets ($\theta_{\text{j}} > 0.1$ rad) do not reproduce the rapid decay after the peak in the light curve. In a different study²⁷, we carried out a full scan of the parameter space using two different semi-analytical jet structures, and the values obtained for θ_{j} and θ_{obs} lie within the ranges specified above.

Our simulation that provides the best fit to the data is of a 0.08 -rad (4° at the time of light-curve peak) jet that is observed from $\theta_{\text{obs}} = 0.35$ rad (20°). In this simulation, the cocoon dominates the observed radio emission until about day 60, after which time the jet dominates (see Fig. 2 and Methods). The Lorentz factor of the observed region decreases slowly from $\Gamma \approx 4$ on day 75 to $\Gamma \approx 3$ on day 230. Within the framework of standard afterglow theory from a successful jet, the observations put tight constraints on additional properties of the jet and the surrounding environment (see Methods). The total energy of the relativistic ejecta (jet + cocoon) is in the range 10^{49} – 10^{50} erg and the external density is 10^{-4} cm^{-3} to $5 \times 10^{-3} \text{ cm}^{-3}$. Figure 4 illustrates the physical and geometric parameters that we derive for GW170817.

Our final model is qualitatively similar to jet + cocoon (also referred to as structured jet) models suggested previously^{13,15,16,28}. However, owing to the VLBI data and more up-to-date light curves, our constraints on the opening and viewing angles of the jet are much tighter than those obtained from previous models, and in tension with some. The small viewing angle (around 20°) for GW170817 is expected in only about 5% of the mergers (not accounting for the gravitational-wave polarization bias). Our best-fitting model suggests we were relatively lucky with GW170817 because the afterglow of this event as observed at larger angles would be much fainter. In our best-fitting numerical model, the radio emission should be detectable at a viewing angle of about 30° , but would probably be too faint for detection at an angle of about 40° . The detectability of future GW170817-like events depends on the circum-merger density. Taking our best-fitting model for GW170817, but increasing the density to 0.01 cm^{-3} (the median

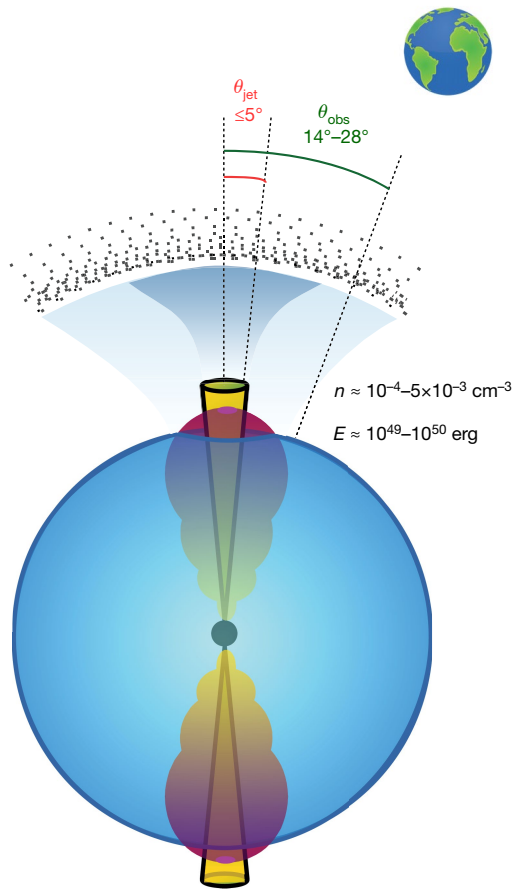


Fig. 4 | Schematic of the physical and geometric parameters derived for GW170817. GW170817 has a successful jet (yellow) that drives a cocoon (red) through interaction with the dynamical ejecta (blue). This scenario is the same as scenario E in our previous work⁸ and consistent with structured-jet models. The shock breakout from the cocoon probably produced the γ -ray signal and the cocoon's interaction with the interstellar medium produced the early-time (up to about two months after the merger) radio and X-ray emission. The relativistic core of the jet has a half-opening angle of $\theta_{\text{jet}} \leq 5^\circ$. Earth is located $\theta_{\text{obs}} = 14^\circ\text{--}28^\circ$ away from the core of the jet. GW170817 most probably gave rise to a SGRB pointing at such an angle away from Earth. The interaction between the jet and the interstellar medium produced the late-time radio and X-ray emission. Our VLBI measurement suggests that the Lorentz factor of the jet 150 days after the merger (at the peak of the radio light curve, when the core of the jet came into view) is $\Gamma \approx 4$. The total energy (E) of the jet and cocoon system is $10^{49}\text{--}10^{50}$ erg. The density (n) of the circum-merger environment is $10^{-4}\text{--}5 \times 10^{-3} \text{ cm}^{-3}$.

density²⁹ for short GRBs (SGRBs); while keeping all other values constant), we find an afterglow that is brighter by about an order of magnitude at the peak compared to that of GW170817. Such an afterglow could have been detected at a distance of 40 Mpc at a larger viewing angle of about 50° .

Our VLBI result implies that binary neutron-star mergers launch relativistic, narrowly collimated jets that successfully penetrate the dynamical ejecta, which is a prerequisite for the production of SGRBs (which require $\Gamma_0 \gtrsim 100$). If GW170817 produced an SGRB pointing away from us, then its peak isotropic equivalent luminosity in γ -rays was $L_{\text{iso}} \approx 10^{52} \text{ erg s}^{-1}$ when observed within the jet cone, assuming that the initial opening angle of the jet was around 0.05 rad. The rate of SGRBs with a peak L_{iso} of more than about $10^{52} \text{ erg s}^{-1}$ is only³⁰ about $0.1 \text{ Gpc}^{-3} \text{ yr}^{-1}$, corresponding to about 1% of all SGRBs that point towards Earth. This suggests either that we were extremely lucky to observe such an event or that all such luminous events are more narrowly beamed than events of smaller L_{iso} and do not typically point towards Earth. For example, if GW170817, with an opening angle of

approximately 0.05 rad, is representative of events with $L_{\text{iso}} \approx 10^{52} \text{ erg s}^{-1}$, it would imply that there are 1,000 events with such luminosity that point away from Earth for every SGRB-producing event that points towards Earth—that is, a rate of about $100 \text{ Gpc}^{-3} \text{ yr}^{-1}$ for GW170817-like events. This rate is about 3%–30% of the binary neutron-star merger rate¹ ($1,540^{+3,200}_{-1,220} \text{ Gpc}^{-3} \text{ yr}^{-1}$) and would imply that the true fraction of high-luminosity SGRBs is much higher than observed at Earth. An anticorrelation between the opening angle of the jet and its isotropic equivalent energy is one possible cause for such a relationship, and follows naturally if the total energy of different events varies less than their beaming. This possibility can be easily tested with a small number of future events with off-axis afterglow emission.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0486-3>.

Received: 19 June; Accepted: 7 August 2018;

Published online 5 September 2018.

- Abbott, B. P. et al. GW170817: observation of gravitational waves from a binary neutron star inspiral. *Phys. Rev. Lett.* **119**, 161101 (2017).
- Abbott, B. P. et al. Multi-messenger observations of a binary neutron star merger. *Astrophys. J.* **848**, L12 (2017).
- Hjorth, J. et al. The distance to NGC 4993: the host galaxy of the gravitational-wave event GW170817. *Astrophys. J.* **848**, L31 (2017).
- Hallinan, G. et al. A radio counterpart to a neutron star merger. *Science* **358**, 1579–1583 (2017).
- Troja, E. et al. The X-ray counterpart to the gravitational-wave event GW170817. *Nature* **551**, 71–74 (2017).
- Margutti, R. et al. The electromagnetic counterpart of the binary neutron star merger LIGO/Virgo GW170817. V. Rising X-ray emission from an off-axis jet. *Astrophys. J.* **848**, L20 (2017).
- Haggard, D. et al. A deep Chandra X-ray study of neutron star coalescence GW170817. *Astrophys. J.* **848**, L25 (2017).
- Moolley, K. P. et al. A mildly relativistic wide-angle outflow in the neutron-star merger event GW170817. *Nature* **554**, 207–210 (2018).
- Dobie, D. et al. A turnover in the radio light curve of GW170817. *Astrophys. J.* **858**, L15 (2018).
- Alexander, K. D. et al. A decline in the X-ray through radio emission from GW170817 continues to support an off-axis structured jet. *Astrophys. J.* **863**, L18 (2018).
- Kasliwal, M. M. et al. Illuminating gravitational waves: A concordant picture of photons from a neutron star merger. *Science* **358**, 1559–1565 (2017).
- Troja, E. et al. The outflow structure of GW170817 from late-time broad-band observations. *Mon. Not. R. Astron. Soc.* **478**, L18–L23 (2018).
- Xie, X., Zrake, J. & MacFadyen, A. Numerical simulations of the jet dynamics and synchrotron radiation of binary neutron star merger event GW170817/GRB 170817A. *Astrophys. J.* **863**, 58 (2018).
- Lamb, G. P. & Kobayashi, S. Electromagnetic counterparts to structured jets from gravitational wave detected mergers. *Mon. Not. R. Astron. Soc.* **472**, 4953–4964 (2017).
- Lazzati, D. et al. Late time afterglow observations reveal a collimated relativistic jet in the ejecta of the binary neutron star merger GW170817. *Phys. Rev. Lett.* **120**, 241103 (2018).
- Margutti, R. et al. The binary neutron star event LIGO/Virgo GW170817 160 days after merger: synchrotron emission across the electromagnetic spectrum. *Astrophys. J.* **856**, L18 (2018).
- Lyman, J. D. et al. The optical afterglow of the short gamma-ray burst associated with GW170817. *Nat. Astron.* <https://doi.org/10.1038/s41550-018-0511-3> (2018).
- Resmi, L. et al. Low frequency view of GW 170817/GRB 170817A with the Giant Meterwave Radio Telescope. Preprint at <https://arxiv.org/abs/1803.02768> (2018).
- Nakar, E. & Piran, T. Implications of the radio and X-ray emission that followed GW170817. *Mon. Not. R. Astron. Soc.* **478**, 407–415 (2018).
- Troja, E., Piro, L. & Ryan, G. Chandra observations of GW170817 reveal a fading afterglow. *Astron. Teleg.* 11619 (2018).
- Hotkezaka, K., Kiuchi, K., Shibata, M., Nakar, E. & Piran, T. Synchrotron radiation from the fast tail of dynamical ejecta of neutron star mergers. Preprint at <https://arxiv.org/abs/1803.00599> (2018).
- D'Avanzo, P. et al. The evolution of the X-ray afterglow emission of GW 170817/GRB 170817A in XMM-Newton observations. *Astron. Astrophys.* **613**, L1 (2018).
- Gill, R. & Granot, J. Afterglow imaging and polarization of misaligned structured GRB jets and cocoons: breaking the degeneracy in GRB 170817A. *Mon. Not. R. Astron. Soc.* **478**, 4128–4141 (2018).
- Taylor, G., Frail, D., Berger, E. & Kulkarni, S. High resolution observations of GRB 030329. *AIP Conf. Ser.* **727**, 324–327 (2004).
- Boutelier, T., Henri, G. & Petrucci, P.-O. The influence of the jet opening angle on the appearance of relativistic jets. *Mon. Not. R. Astron. Soc.* **418**, 1913–1922 (2011).

26. Lind, K. R. & Blandford, R. D. Semidynamical models of radio jets: relativistic beaming and source counts. *Astrophys. J.* **295**, 358–367 (1985).
27. Hotokezaka, K. et al. A Hubble constant measurement from superluminal motion of the jet in GW170817. Preprint at <https://arxiv.org/abs/1806.10596> (2018).
28. Nakar, E., Gottlieb, O., Piran, T., Kasliwal, M. M. & Hallinan, G. From γ to radio - the electromagnetic counterpart of GW 170817. Preprint at <https://arxiv.org/abs/1803.07595> (2018).
29. Fong, W., Berger, E., Margutti, R. & Zauderer, B. A. A decade of short-duration gamma-ray burst broadband afterglows: energetics, circumburst densities, and jet opening angles. *Astrophys. J.* **815**, 102 (2015).
30. Wanderman, D. & Piran, T. The rate, luminosity function and time delay of non-Collapsar short GRBs. *Mon. Not. R. Astron. Soc.* **448**, 3026–3037 (2015).

Acknowledgements We are grateful to the VLBA, VLA and GBT staff, especially M. Claussen, A. Mioduszewski, T. Minter, F. Ghigo, W. Briskin, K. O'Neill and M. McKinnon, for their support with the HSA observations. We thank V. Dhawan and P. Demorest for help with observational issues with the VLBI system at the VLA. K.P.M. thanks A. Mioduszewski, E. Mornjian, E. Greisen, T. Pearson and S. Kulkarni for discussions. We thank M. Kasliwal for providing comments on the manuscript. The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities. K.P.M. is currently a Jansky Fellow of the National Radio Astronomy Observatory. K.P.M. acknowledges support from the Oxford Centre for Astrophysical Surveys, which is funded through the Hintze Family Charitable Foundation, for some initial work presented here. E.N. acknowledges the support of an ERC starting grant (GRB/SN) and an ISF grant (1277/13).

A.T.D. is the recipient of an Australian Research Council Future Fellowship (FT150100415). G.H. acknowledges the support of NSF award AST-1654815. A.H. acknowledges support by the I-Core Program of the Planning and Budgeting Committee and the Israel Science Foundation. A.C. acknowledges support from the NSF CAREER award number 1455090 titled 'CAREER: Radio and gravitational-wave emission from the largest explosions since the Big Bang'.

Author contributions K.P.M., A.T.D., S.B., G.H. and D.A.F. coordinated the VLBI observations. A.T.D. and K.P.M. performed the VLBI data processing. O.G. and E.N. carried out the theoretical study, including analytic calculations and numerical simulations, with some input from K.H. K.P.M., A.T.D., E.N., G.H. and D.A.F. wrote the paper. A.C. and A.H. compiled the references. A.H., A.D. and K.P.M. compiled Methods. O.G., A.T.D., A.H. and K.P.M. prepared the figures. All co-authors discussed the results and provided comments on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0486-3>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to K.P.M. and O.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Observations, data processing and basic analysis. To establish the size and morphology of the faint radio afterglow of GW170817, we obtained Director's Discretionary Time (programme ID BM469) to observe with the HSA. The HSA antennas included the ten VLBA dishes, the phased Karl G. Jansky VLA and the GBT, although not all stations were used in all observations. The maximum baseline was typically 7,500–8,000 km.

VLBI observations. We observed GW170817 with the HSA over four epochs between 2017 September and 2018 April. Each epoch consisted of 2–4 observations carried out over a period of up to 10 days, with approximately three hours of on-source time on GW170817 per day. The choice of the observing frequency was informed by the results from the VLA monitoring of the radio light curve, the desired angular resolution and the ease of scheduling on the telescopes. In all epochs, a total bandwidth of 256 MHz was sampled in dual polarization at 2-bit precision. Depending on the observing frequency, the recorded bandwidth was broken into eight 32-MHz-wide bands or two 128-MHz-wide bands. A summary of the observations is given in Extended Data Table 1.

The first epoch was undertaken in the L band (central frequency of 1,550 MHz) 37–38 days post-merger. No fringes were seen on the GBT on one of the two observing days owing to an unknown technical issue, considerably reducing overall sensitivity at this epoch. The second epoch was carried out in the S band (central frequency of 3,200 MHz), 51–52 days post-merger. However, a misconfiguration of the VLA correlator on both days meant that phased VLA data were practically unusable, and hence sensitivity was severely impacted. The third epoch was observed in the C band (central frequency of 4,540 MHz) 72–79 days post-merger. The fourth epoch was likewise observed in the C band, 227–236 days post-merger, using only the VLBA and VLA as the GBT was unavailable.

Each observation was structured around an 8-min cycle as follows. We used the source J1258–2219 (an approximately 1-Jy flat-spectrum source, separated by 2.8° from GW170817) as the primary delay and gain calibrator, visiting it twice per cycle during the first three epochs and once per cycle in the fourth epoch of observations. J1312–2350, a 20-mJy source separated by 0.8° from GW170817, was used as a secondary phase calibrator, and was visited once per cycle in the first three epochs and twice per cycle in the fourth epoch of observations. J1258–2219 was additionally used to determine phase solutions for the VLA once per cycle. A single scan on 3C286 was included at the end of each observation to allow flux calibration of the commensally recorded VLA interferometer data. For the C-band (4.5 GHz) epochs only, we included three scans on the blazar OQ208 (B1404+286) over the course of each observation to enable polarization calibration to be determined and applied.

VLBI data processing. We followed standard data-reduction procedures for HSA data using the AIPS software package³¹. For all calibration steps that involve a sky source (fringe-fitting, leakage and self-calibration) we used a model of the source that was iteratively refined over several passes of the entire data-reduction pipeline.

The data was loaded using 'FITLD' and a priori amplitude corrections were applied using 'ANTAB' and 'ACCOR'. An issue with the VLA automatic gain control was uncovered whereby the phased VLA data exhibited large short-term amplitude variations; this could be (and was) largely mitigated by using a per-integration solution for the auto-correlation-based corrections with 'ACCOR', but small residual variations that were weakly detrimental to sensitivity remained. This problem was fixed before the fourth observational epoch. 'CLCOR' was used to correct for parallactic angle rotation and to apply the most accurate available values for Earth orientation parameters. 'TECOR' was used to correct for ionospheric propagation effects, using the 'igsg' model available from <ftp://cddis.gsfc.nasa.gov/gps/products/ionex>. We then calibrated the time-independent delays and the antenna bandpass using 'FRING' and 'BPASS'; in the first two epochs we used a scan on the primary calibrator J1258–2219, whereas in the third and fourth epochs we used OQ208.

For the third epoch at 4.5 GHz only, we calibrated the cross-polar delays and instrumental polarization leakage using the tasks 'FRING' and 'LPCAL' and the source OQ208. This step was essential because of the large (roughly 30%) leakage at the GBT at this frequency. 'LPCAL' solves for a single leakage value per sub-band, whereas the GBT polarization leakage varies across the 128-MHz subband; accordingly, we split each 128-MHz subband into four 32-MHz subbands to allow a coarse frequency dependence to the leakage solutions.

We solved for time-dependent delays using 'FRING' on the primary gain calibrator J1258–2219, followed by self-calibration on this source using 'CALIB', obtaining a single solution per subband per scan. Finally, we improved the phase calibration using self-calibration on the secondary gain calibrator J1312–2350, deriving a single frequency-independent solution per scan.

At each stage, the solutions from the SN table were applied to the CL table using 'CLCAL'. The final CL table was applied to the target using 'SPLIT'. The target was then exported in UVFITS format using 'FITP' and imaged using 'difmap'³². **VLA/VLBI interferometric data processing.** We processed using VLA cross-correlated data (with the WIDAR correlator) using a custom-developed pipeline,

which incorporates manual flagging, and standard interferometric data-calibration techniques in CASA. The imaging was done with the CASA task clean with natural weighting, choosing an image size of $4,096 \text{ pixels} \times 4,096 \text{ pixels}$ and a cell size of 0.5 arcsec .

The VLA-only data give the GW170817 flux densities of $56 \pm 8 \mu\text{Jy beam}^{-1}$, $54 \pm 8 \mu\text{Jy beam}^{-1}$ and $45 \pm 7 \mu\text{Jy beam}^{-1}$ for the three observations of the third epoch at 4.5 GHz. All three observations combined give $55 \pm 5 \mu\text{Jy beam}^{-1}$. For the four observations of the fourth epoch, the flux density values are $55 \pm 8 \mu\text{Jy beam}^{-1}$, $46 \pm 8 \mu\text{Jy beam}^{-1}$, $48 \pm 6 \mu\text{Jy beam}^{-1}$ and $46 \pm 6 \mu\text{Jy beam}^{-1}$; all four observations combined give $48 \pm 4 \mu\text{Jy beam}^{-1}$.

Flux comparison between the VLBI and VLA interferometric data. A comparison between the flux densities measured in the VLA-only interferometric data and those measured in the VLBI data (see Extended Data Table 1) implies that, within 1σ uncertainties (typically 10% of the source flux density), no flux is being resolved out in the VLBI data.

Model fits and parameter estimates. Difmap³² was first used to produce a 'dirty' (un-deconvolved) image from the concatenated data from each epoch and the individual observations within each epoch. In the first two epochs, there was substantial loss of sensitivity owing to technical issues, and the source was not detected. We place 5σ upper limits of $40 \mu\text{Jy beam}^{-1}$ (1.6 GHz, day 38) and $60 \mu\text{Jy beam}^{-1}$ (3.2 GHz, day 52) on the flux densities of GW170817, and do not consider these epochs further.

In the third and fourth epochs, a radio counterpart to GW170817 can clearly be seen in the dirty images for the concatenated datasets, and the source can also be seen (albeit at low signal-to-noise ratio) in the individual observations. Initially, we fit the data in the visibility plane using a single circularly symmetric Gaussian model component. Although probably an over-simplification of the true source structure, this has the advantage of being fast and simple to fit, while providing an accurate estimate of the flux centroid position. After model fitting, we read the resultant clean image into AIPS and used the task JMFIT to fit an elliptical Gaussian in the image plane. Compared to model fitting, this has the advantage of providing well-constrained estimates of the uncertainty of the key parameters of interest³³. In the third epoch (day 75), the best-fit values of flux density and position are $58 \pm 5 \mu\text{Jy beam}^{-1}$, RA = 13 h 09 min 48.068638(9) s and dec. = $-23^\circ 22' 53.3909(4)''$. The uncertainties given here are purely statistical; we consider systematic contributions in the following sections. The best-fit size was a full-width at half-maximum (FWHM) of 0.0 mas ; that is, the source was modelled as a point source. At day 230, the best-fit values of flux density and position are $48 \pm 6 \mu\text{Jy beam}^{-1}$, RA = 13 h 09 min 48.068831(11) s and dec. = $-23^\circ 22' 53.3907(4)''$, and the best-fit de-convolved size is 0.7 mas , although an unresolved source could not be excluded. The images of the source at days 75 and 230 are shown in Extended Data Fig. 1.

Estimating systematic contributions to flux density and position uncertainties. The absolute calibration of flux densities in VLBI maps can be challenging owing to the fact the sources compact enough to be visible at milliarcsecond resolution typically evolve on a timescale of months to years. In cases where only a priori amplitude calibration can be performed, the accuracy of the flux density scale of a VLBI image is typically assumed to be roughly 20%. In this case, we are able to use the contemporaneous VLA data to establish an absolute flux density scale, using the calibrator sources J1312–2350 and J1258–2219 (under the assumption that these sources do not have substantial structure on scales larger than that resolvable by our VLBI observations). After adjusting the VLBI amplitude scale to produce the closest match to these two sources, the residual differences are typically 10% for each observation, and hence systematic uncertainties on our measured values of flux density for GW170817 are comparable to our statistical uncertainties.

Similarly, for our image centroid positions, we must consider the possibility of systematic position shifts between epochs owing to calibration errors, in addition to the limiting precision attainable on the basis of the image resolution and signal-to-noise ratio. We neglect systematic errors due to the uncertainty in the calibrator reference position, because this would affect both epochs equally. Given the relatively close proximity of our calibrator source J1312–2350 to GW170817 (0.8°), we expect any systematic errors that vary between epochs to be at most a small fraction of the synthesized beam size. Astrometric simulations³⁴ suggest a typical systematic error for a single observation with the VLBA of 0.07 mas in RA and 0.25 mas in dec. for our observing conditions (dec. = -26° , angular separation of 0.8°). However, these simulations do not include the effect of the ionosphere, which could treble the systematic error at an observing frequency of 4.5 GHz under typical conditions. Countering this, our epochs consist of 3–4 observations spread over about 7 days, and systematic errors (in particular those due to the ionosphere) are likely to be only weakly correlated over this timescale. On the basis of these considerations, we estimated the systematic position uncertainty to be 0.15 mas in RA and 0.5 mas in dec., and added this value in quadrature with the formal position fit errors at each epoch.

To verify this expectation, we repeated the data reduction for the third and fourth epochs after shifting the phase centre of our target field to the position of

the NGC 4993 low-luminosity active galactic nucleus. This source is separated by 10.3 arcsec from GW170817, and hence falls outside the field of view of the phased VLA; accordingly, the VLA was flagged before imaging. The positions obtained for the active galactic nucleus have a separation of 0.05 mas in RA and 0.5 mas in dec. (see Extended Data Fig. 2). This is consistent with both their statistical uncertainties and our estimate for the systematic errors derived above. The flux density of the active galactic nucleus is consistent with a constant value (0.25 ± 0.02 mJy and 0.29 ± 0.03 mJy in the third and fourth epochs, respectively, where the 1σ uncertainties are purely statistical).

Comparison between the VLBI data and synthetic images. To compare the models with our VLBI data, we converted the simulated images (example images shown in Fig. 3; for details of the simulations see the next section) into difmap models consisting of point sources at the centre of each non-zero pixel in the simulated image, and performed model fitting in the visibility plane. The rotation, translation and total flux density of the image were taken as free parameters, although we used the approximate positions and flux densities from our earlier fitting of circular Gaussian components to restrict the ranges of parameter values over which we searched. For each model, we recorded the χ^2 obtained at the best-fit values for rotation, translation and total flux density.

Because the signal-to-noise ratio of each individual visibility measurement is very low, determining the increase in χ^2 that indicates a significant discrepancy between models is not straightforward. Previous authors have often relied on visual inspection of images and visibility data to determine model goodness-of-fit^{35,36}. Owing to the low signal-to-noise ratio of our target image, we took a different approach. First, we used an image-plane fit to determine the position errors in the image plane using the dataset fitted with a circular Gaussian component, which is a well-understood process³³. Second, we perturbed the position of the circular Gaussian model component by up to $\pm 3\sigma$ in RA and $\pm 3\sigma$ in dec., and recorded the change in χ^2 at offsets of 1σ , 2σ and 3σ . A consistent increase in χ^2 was seen regardless of the direction of the positional perturbation. Finally, we fitted other models based on the hydrodynamic simulations to the data and recorded the χ^2 in each case. The reference positions for a given model were allowed to vary between the day-75 and day-230 datasets by up to the amount of our estimated systematic position uncertainty of 0.15 mas in RA and 0.5 mas in dec. By comparison to the set of χ^2 values obtained from the perturbed circular Gaussian fits, we estimated the consistency of each hydrodynamic model with the best-fitting circular Gaussian model.

In addition to fitting the actual synthetic images, we first produced an estimate of the maximum source extent, by finding the largest circular and elliptical Gaussian sources that produced a χ^2 that did not deviate by more than 1σ from the best circular Gaussian fits. For the epochs at day 75 and day 230, the largest circular Gaussian source was 1.1 mas and 1.2 mas in diameter, respectively. The best-fitting elliptical Gaussian converged to an unphysical one-dimensional source for each epoch, with an upper limit on the major axis of 12 mas and 9 mas for day 75 and day 230, respectively. In both cases the best-fit position angle was approximately aligned with the major axis of the beam and hence approximately perpendicular to direction of source motion. Tighter limits on the maximum size can be obtained if the axial ratio of the elliptical Gaussian source is constrained to a physical value: for instance, in the case of the day-230 dataset, the largest source permitted with an axial ratio of 4:1 is $3.9 \text{ mas} \times 0.9 \text{ mas}$. Hence, the source size parallel to the direction of motion is relatively well constrained.

None of the synthetic images produced a χ^2 significantly better than a simple circular Gaussian in either epoch (unsurprising, given that the source was consistent with being unresolved in both cases). Generally, we found that as the positional offset between days 75 and 230 increased, the best-fit source size at day 230 also increased and was often inconsistent with the observed compactness of the source. This disfavoured models at low viewing angles. Conversely, models at large viewing angles were incapable of producing a sufficiently large positional offset.

The best-fitting model (narrow jet viewed at 0.35 rad, model A1 in Figs. 2 and 3) was able to produce the expected positional shift between epochs: with a constant reference translation and rotation, it produced an acceptable fit to both the day-75 epoch (χ^2 increase equivalent to a 0.9σ position offset for the circular Gaussian) and the day-230 epoch (χ^2 increase equivalent to a 1.3σ position offset for the circular Gaussian). Of the other models, only one (model B, the very narrow jet viewed at 0.3 rad) remained consistent within 2σ for both epochs. For all other models, the discrepancy with the best-fitting circular Gaussian exceeded 2σ in one or both epochs. As can be seen in Fig. 2, models A1 and B are also those that best fit the light curve.

Numerical hydrodynamic simulations. To characterize the properties of different models, we carry out relativistic hydrodynamical simulations of various set-ups, followed by a post-processing numerical calculation²⁸ of their afterglow light curve and observed images at 75 and 230 days. In particular, we run different types of model to see which have the potential to fit the entire dataset of both the light curve and the image characteristics, that is, the flux centroid movement and the image size constraints.

Our set-up includes three components: the jet, a core of cold massive ejecta and a fast ejecta tail. Each component of the ejecta expands homologously and has a density profile of

$$\rho(r, \theta) = \rho_0 r^{-\alpha} \left(\frac{1}{4} + \sin^2 \theta \right)$$

where the normalization ρ_0 is determined by the total ejecta mass, and α and β , which differ between models, dictate the radial and angular structures, respectively. However, our main focus was on scanning the properties of jets, such as luminosities, opening angles, injection and delay times. Although some of the jets successfully break out from the ejecta if their properties allow, others may be choked inside it. We ran about ten different models; here we present four representative models that demonstrate how the different characteristics of the jet affect the observed outcome. The first two models are narrow jets and are found to fit all of the observed characteristics—the gradual rise of the flux, the short plateau at the peak followed by a fast decline and the large flux centroid motion between the two image epochs. In addition, we present a wider successful jet and a choked jet. The full set-up is given in Extended Data Table 2.

A full description of the hydrodynamic simulations is given in our previous work²⁸. In brief, for each model we use three different simulations. The first one, which includes the jet propagation inside the core ejecta, is performed in three dimensions to avoid the numerical plug artefact³⁷. The second simulation includes the outflow evolution inside the tail ejecta and after breaking out of it until reaching the homologous phase. This simulation is modelled in two dimensions because after breakout the plug artefact is no longer a concern³⁸, and two- and three-dimensional simulations become similar. Finally, the third simulation begins when the afterglow becomes important and ends after it decays.

For the relativistic hydrodynamical simulation we use the public code PLUTO³⁹ v4.0 with an Harten–Lax–van Leer Riemann solver and apply an equation of state with an adiabatic index of 4/3. The set-up of models A and B is as follows. The grid set-up of the first three-dimensional Cartesian simulation has three patches on the x and y axes and two patches on the z axis. On x and y the inner patch spans from -2×10^8 cm to 2×10^8 cm with 30 uniform cells. The outer patch is from $[2 \times 10^8 \text{ cm}]$ to $[3 \times 10^{10} \text{ cm}]$ with 400 cells that are distributed logarithmically. On the z axis the first patch is uniform from 4.5×10^8 cm to 10^{10} cm with 200 cells, followed by a logarithmic patch of 400 cells until 4×10^{10} cm. We convert the three-dimensional output of the first simulation to an axisymmetric grid³⁸, which is the initial set-up of the second simulation, for which the set-up is as follows. The first two patches on the r and z axes correspond to the three-dimensional set-up. We add another patch on each axis from 3×10^{10} cm (4×10^{10} cm) on the r (z) axis to 6×10^{11} cm, with 1,200 logarithmic cells.

For the third simulation, which includes two patches on each axis, we use the output of the second simulation. The first patch corresponds to the second simulation grid with 800 uniform cells until $6 \times 10^{11} \times R$ cm on each axis. The second patch on each axis stretches to $10^{14} \times R$ cm with 6,000 logarithmic cells. Because the simulation is dimensionless, we use R as a scaling length factor²⁸; R also determines the density of the interstellar medium (ISM), which is set to be $\rho_{\text{ISM}} = 5 \times 10^{-12} \text{ g} \times (R \times \text{cm})^{-3}$ in simulation A and $\rho_{\text{ISM}} = 8 \times 10^{-12} \text{ g} \times (R \times \text{cm})^{-3}$ in simulation B. Each viewing-angle fit requires a different R . The best fits for $\theta_{\text{obs}} = 0.25, 0.35$ and 0.45 in simulation A are obtained at $R = 3 \times 10^5, 1.7 \times 10^5$ and 8.3×10^4 , respectively; for $\theta_{\text{obs}} = 0.3$ in simulation B it is $R = 5 \times 10^5$.

The set-up of simulations C and D has been described previously²⁸ (simulation D is identical to the successful-jet scenario, except for the engine time), and the only difference here is that for the outer patch in the third part we use a high resolution of 4,000 cells rather than the 2,500 cells used originally. The scaling of the third part of the simulation is determined by $n = 4 \times 10^{-2} \text{ cm}^{-3}$ and $n = 4.5 \times 10^{-3} \text{ cm}^{-3}$ in C and D, respectively.

Finally, we verify that each of the three simulations meets the required resolution to reach convergence. We first compare the resolution of the first two simulations, from the jet launch until reaching the homologous phase, with previously published simulations³⁷ for which convergence tests have been done. The resolution of the three-dimensional simulation that handles the jet propagation inside the ejecta is comparable with that of the inner parts of the previous simulations. The sequential two-dimensional simulation naturally has a higher resolution compared with the outer parts of the three-dimensional grid presented previously³⁷. For convergence of the third part in which the outflow interacts with the ISM, we perform another set of simulations with 2/3 of the aforementioned resolution. We find that both the light curves and the images for the relevant viewing angles remain essentially unchanged with the increase in resolution.

Details of the simulation that provides the best fit to the data. Our simulation that provides the best fit to the data is of a jet with a 0.08-rad (4°) opening angle, at the time of light curve peak, that is observed at a viewing angle of $\theta_{\text{obs}} = 0.35$ rad (20°). In this simulation, a relativistic jet is injected into the sub-relativistic merger

ejecta. The jet is followed during its propagation through the ejecta, the formation of the cocoon and the breakout of the jet and the cocoon from the dynamical (sub-relativistic) ejecta. The simulation then continues to follow the interaction of the outflow (jet + cocoon) with the ISM. When this interaction starts, the opening angle of the jet is 0.04 rad. The cocoon dominates the observed radio emission during the first approximately 60 days, after which time the jet dominates. The jet expands sideways slowly during its interaction with the ISM, reaching an opening angle of 0.08 rad after about 150 days at the light-curve peak. On day 75, the Lorentz factor of the observed region is $\Gamma \approx 4$, which steadily drops to $\Gamma \approx 3$ by day 230.

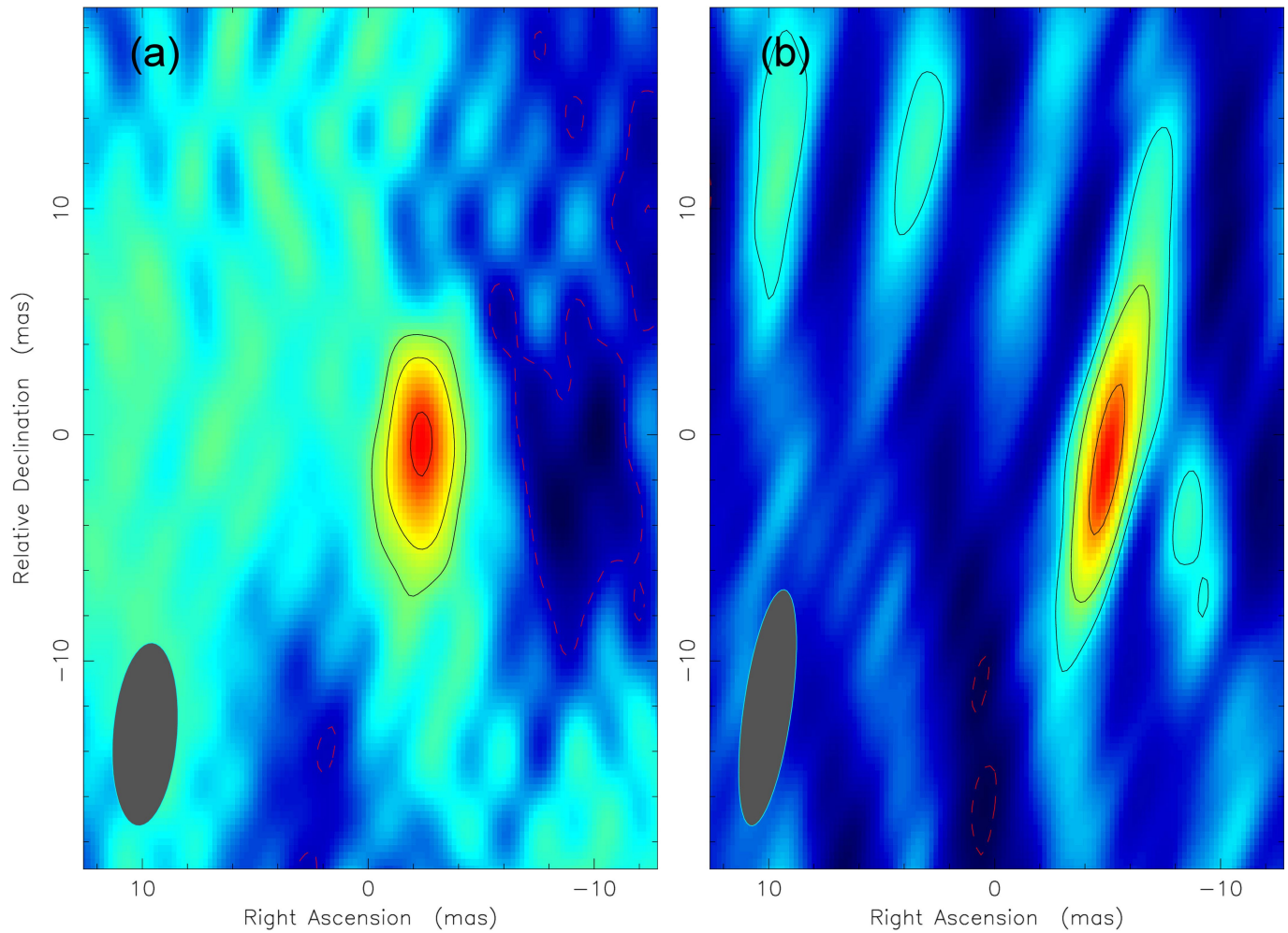
Constraining the jet energy and the external density. The γ -ray signal from GW170817 had an isotropic equivalent energy of 5×10^{47} erg. The afterglow suggests that this energy is not representative of the jet energy. This is consistent with models for the γ -ray emission^{11,38,40–44}. Therefore, to constrain the jet energy and external density, we use the constraints on the geometry of the outflow together with the observed afterglow light curve to constrain the outflow energy. We use the standard afterglow model, where a narrow ultra-relativistic jet drives a blast wave into the external medium, which radiates in synchrotron emission to produce the radio and X-ray afterglow. Before interacting with the external medium, the jet has an initial Lorentz factor Γ_0 . This is also the initial Lorentz factor of the blast wave that it drives, which is constant at first, until the blast wave accumulates enough mass and starts decelerating. The initial opening angle of the jet $\theta_{j,0}$ is also constant until the Lorentz factor drops to about $1/\theta_{j,0}$. At this point, if $\theta_{j,0} < 0.05$ rad the jet starts spreading sideways rapidly until $\theta_{j,0} \approx 0.05$ rad, at which point it starts spreading sideways more slowly⁴⁵. We have direct constraints on Γ and θ_j only near the time of the peak of the light curve. We can therefore put only a lower limit on the initial Lorentz factor, $\Gamma_0 > 4$, and an upper limit on the initial opening angle, $\theta_{j,0} < 0.1$ rad. Moreover, given the fast spreading of the jet if $\theta_{j,0} < 0.1$ rad and $\Gamma < 1/\theta_j$, at the time that we observe the jet its opening angle is expected to be $\theta_j \approx 0.05$ – 0.1 rad even if initially $\theta_{j,0} \ll 0.1$ rad and $\Gamma_0 \gg 4$. The Lorentz factor and the time of the peak provide a relation between the density of the ambient medium (assumed to be constant) and the isotropic equivalent energy of the jet¹⁹: $E_{\text{iso}} \approx 10^{52} n / (3 \times 10^{-4} \text{ cm}^{-3})$ erg. The flux is extremely sensitive to the Lorentz factor and we can use its value at the peak to constrain the density and the fraction of the internal energy that goes to the magnetic field¹⁹, $\varepsilon_B: n / (3 \times 10^{-4} \text{ cm}^{-3}) \times (\varepsilon_B / 10^{-3})^{0.47} \approx (\Gamma / 3.5)^{5.9}$, where we assume that 10% of the internal energy goes to the accelerated electrons ($\varepsilon_e = 0.1$) and that their distribution power-law index is $p = 2.16$. By allowing the least-constrained parameter, ε_B , to vary between 10^{-2} and 10^{-5} , we find that the circum-merger density is 10^{-4} – $5 \times 10^{-3} \text{ cm}^{-3}$ and that the jet isotropic equivalent energy is $E_{\text{iso}} \approx 3 \times 10^{51}$ – 10^{53} erg. Because the opening angle of the jet at this time is 0.05–0.1 rad and because the jet contains a substantial fraction of the total energy of the relativistic outflow (jet + cocoon), we find that the energy deposited by the merger in relativistic ejecta is 10^{49} – 10^{50} erg. The

confirmation of a successful jet in GW170817 also implies high isotropy of the magnetic field⁴⁶.

Code availability. The hydrodynamic simulations were done using the publicly available code PLUTO. Radio data processing software used were AIPS, DIFMAP and CASA.

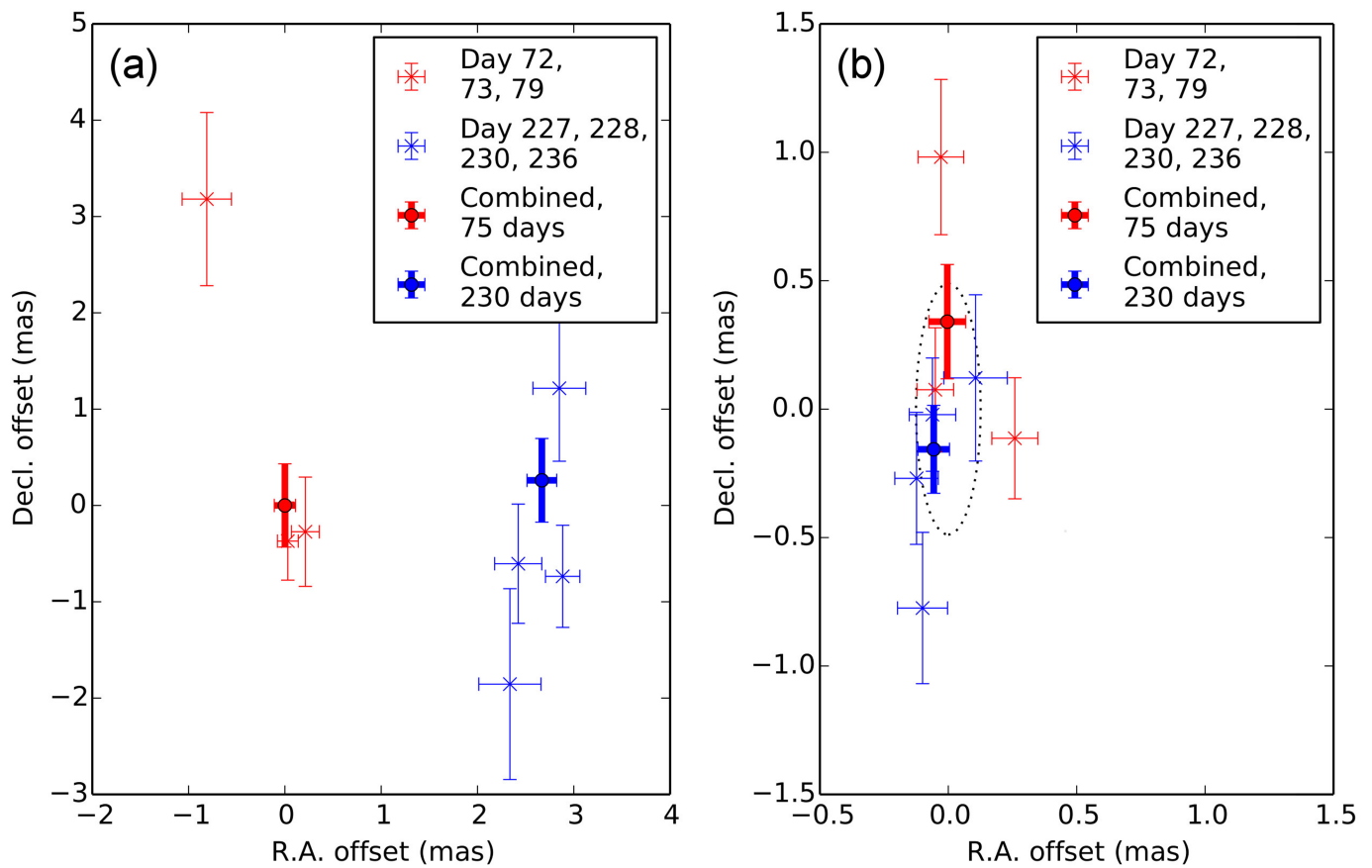
Data availability. All relevant (VLBI) data are available from the corresponding authors on request. The VLA data (presented in Fig. 2) are currently being readied for public release.

31. Greisen, E. W. in *Information Handling in Astronomy – Historical Vistas* (ed. Heck, A.) 109–126 (Kluwer Academic, Dordrecht, 2003).
32. Shepherd, M. C. Difmap: an interactive program for synthesis imaging. *ASP Conf. Ser.* **125**, 77–84 (1997).
33. Condon, J. J. Errors in elliptical Gaussian FITS. *Publ. Astron. Soc. Pacif.* **109**, 166–172 (1997).
34. Pradel, N., Charlot, P. & Lestrade, J.-F. Astrometric accuracy of phase-referenced observations with the VLBA and EVN. *Astron. Astrophys.* **452**, 1099–1106 (2006).
35. Tzioumis, A. K. et al. VLBI observations at 2.3 GHz of the compact galaxy 1934-638. *Astron. J.* **98**, 36–43 (1989).
36. Tingay, S. J., Preston, R. A. & Jauncey, D. L. The subparsec-scale structure and evolution of Centaurus A. II. Continued very long baseline array monitoring. *Astron. J.* **122**, 1697–1706 (2001).
37. Gottlieb, O., Nakar, E. & Piran, T. The cocoon emission – an electromagnetic counterpart to gravitational waves from neutron star mergers. *Mon. Not. R. Astron. Soc.* **473**, 576–584 (2018).
38. Gottlieb, O., Nakar, E., Piran, T. & Hotokezaka, K. A cocoon shock breakout as the origin of the γ -ray emission in GW170817. *Mon. Not. R. Astron. Soc.* **479**, 588–600 (2018).
39. Mignone, A. et al. PLUTO: a numerical code for computational astrophysics. *Astrophys. J. Suppl. Ser.* **170**, 228–242 (2007).
40. Lazzati, D. et al. Off-axis prompt X-ray transients from the cocoon of short gamma-ray bursts. *Astrophys. J.* **848**, L6 (2017).
41. Eichler, D. Testing the viewing angle hypothesis for short GRBs with LIGO events. *Astrophys. J.* **851**, L32 (2017).
42. Kathirgamaraju, A., Barniol Duran, R. & Giannios, D. Off-axis short GRBs from structured jets as counterparts to GW events. *Mon. Not. R. Astron. Soc.* **473**, L121–L125 (2018).
43. Bromberg, O., Tchekhovskoy, A., Gottlieb, O., Nakar, E. & Piran, T. The γ -rays that accompanied GW170817 and the observational signature of a magnetic jet breaking out of NS merger ejecta. *Mon. Not. R. Astron. Soc.* **475**, 2971–2977 (2018).
44. Pozanenko, A. S. et al. GRB 170817A associated with GW170817: multi-frequency observations and modeling of prompt gamma-ray emission. *Astrophys. J.* **852**, L30 (2018).
45. Granot, J. & Piran, T. On the lateral expansion of gamma-ray burst jets. *Mon. Not. R. Astron. Soc.* **421**, 570–587 (2012).
46. Corsi, A. et al. An upper-limit on the linear polarization fraction of the GW170817 radio continuum. *Astrophys. J.* **861**, L10 (2018).



Extended Data Fig. 1 | VLBI images. a, b, The cleaned images (natural weighting; $0.2 \text{ mas pixel}^{-1}$) from the two epochs of VLBI, 75 days (**a**) and 230 days (**b**) post-merger. The centre coordinates for these images are $\text{RA} = 13 \text{ h } 09 \text{ min } 48.069 \text{ s}$, $\text{dec.} = -23^\circ 22' 53.39''$. The black contours are at 11 , 22 and $44 \mu\text{Jy beam}^{-1}$ in both images (red dashed contour is

$-11 \mu\text{Jy beam}^{-1}$). The peak flux density of the sources is $58 \pm 5 \mu\text{Jy beam}^{-1}$ (**a**) and $48 \pm 6 \mu\text{Jy beam}^{-1}$ (**b**) (image root-mean-square noise quoted as the 1σ uncertainty). The ellipse in the lower left corner of each panel shows the synthesized beam: $(12.4, 2.2, -7)$ and $(9.1, 3.2, -4)$ for the two epochs (major axis in mas, minor axis in mas, position angle in degrees).



Extended Data Fig. 2 | VLBI astrometric accuracy. **a, b,** The VLBI positions of GW170817 (**a**, relative to the best-fit position at day 75) and the low-luminosity active galactic nucleus in NGC 4993 (**b**, relative to the previously derived position using VLBA-only observations). The individual observations of GW170817 have very low signal-to-noise ratio and hence large errors; the moderately discrepant measurement on day 72 has the lowest signal-to-noise ratio and was affected by observing issues at

the GBT. The NGC 4993 positions do not show any significant systematic position shifts between the two epochs, and are consistent with our estimated systematic position uncertainties of 0.15 mas in RA and 0.5 mas in dec. The root-mean-square variation in the position of the nucleus of NGC 4993 over our seven individual observations (0.14 mas in RA and 0.49 mas in dec.) is shown as a dotted ellipse in **b**. All error bars and uncertainties quoted are 1σ .

Extended Data Table 1 | Log of VLBI (HSA) observations

Epoch	Date	Time	ν_c	BW	Δt	F_ν	Comments
	(UT)	(UT)	(GHz)	(MHz)	(days)	($\mu\text{Jy/beam}$)	
1	2017 Sep 23	16.5h–22.5h	1.6	256	37	<40	No fringes on the GBT
	2017 Sep 24	16.5h–22.5h			38		
2	2017 Oct 07	15.5h–21.5h	3.2	128	51	<60	VLA mis-configured
	2017 Oct 08	15.5h–18.8h			52		VLA mis-configured
3	2017 Oct 28	14.5h–20.5h	4.5	256	72	58 ± 5	
	2017 Oct 29	14.5h–20.5h			73		
	2017 Nov 04	14.0h–20.0h			79		
4	2018 Apr 01	04.5h–10.5h	4.5	256	227	48 ± 6	VLBA+VLA
	2018 Apr 02	04.5h–10.5h			228		VLBA+VLA
	2018 Apr 04	04.5h–10.5h			230		VLBA+VLA
	2018 Apr 10	04.5h–10.5h			236		VLBA+VLA

ν_c is the centre observing frequency, BW is the effective bandwidth after radio-frequency-interference excision, Δt is the time post-merger and F_ν is the peak flux density of GW170817 (image root-mean-square noise quoted as the 1σ uncertainty; upper limits are 5σ).

Extended Data Table 2 | The initial set-ups of models A–D

Model type	Narrow jets		Wider jet	Choked jet
Model	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
L_j (10^{50} erg)	1.4	0.6	6.7	
θ_{inj}	0.07	0.04	0.18	
t_{inj} (s)	0.2	0.3	0.72	
t_{eng} (s)	0.8	0.6	1.0	0.4
h_j	200	400	80	
M_c ($0.01 M_\odot$)	4		5	
M_t ($10^{-3} M_\odot$)	1.6		2.0	
α_c	2		3.5	
α_t	14		10	
β	8		3	
$v_{max,c}/c$	0.2		0.2	
$v_{max,t}/c$	0.6		0.8	

The parameters of the jet are the total luminosity L_j , opening angle upon injection θ_{inj} , injection delay time since the merger t_{inj} , working engine time t_{eng} and specific enthalpy h_j . The ejecta parameters are its mass M , density radial power law $-\alpha$, density angular distribution β and front velocity v_{max} . Each is given for the core with subscript ‘c’ and tail with subscript ‘t’.

A dynamically young and perturbed Milky Way disk

T. Antoja^{1*}, A. Helmi², M. Romero-Gómez¹, D. Katz³, C. Babusiaux^{3,4}, R. Drimmel⁵, D. W. Evans⁶, F. Figueras¹, E. Poggio^{5,7}, C. Reylé⁸, A. C. Robin⁸, G. Seabroke⁹ & C. Soubiran¹⁰

The evolution of the Milky Way disk, which contains most of the stars in the Galaxy, is affected by several phenomena. For example, the bar and the spiral arms of the Milky Way induce radial migration of stars¹ and can trap or scatter stars close to orbital resonances². External perturbations from satellite galaxies can also have a role, causing dynamical heating of the Galaxy³, ring-like structures in the disk⁴ and correlations between different components of the stellar velocity⁵. These perturbations can also cause ‘phase wrapping’ signatures in the disk^{6–9}, such as arched velocity structures in the motions of stars in the Galactic plane. Some manifestations of these dynamical processes have already been detected, including kinematic substructure in samples of nearby stars^{10–12}, density asymmetries and velocities across the Galactic disk that differ from the axisymmetric and equilibrium expectations¹³, especially in the vertical direction^{11,14–16}, and signatures of incomplete phase mixing in the disk^{7,12,17,18}. Here we report an analysis of the motions of six million stars in the Milky Way disk. We show that the phase-space distribution contains different substructures with various morphologies, such as snail shells and ridges, when spatial and velocity coordinates are combined. We infer that the disk must have been perturbed between 300 million and 900 million years ago, consistent with estimates of the previous pericentric passage of the Sagittarius dwarf galaxy. Our findings show that the Galactic disk is dynamically young and that modelling it as time-independent and axisymmetric is incorrect.

Gaia is a European Space Agency (ESA) mission that was designed primarily to investigate the origin, evolution and structure of the Milky Way, and has recently delivered the largest and most precise census of positions, velocities and other stellar properties of more than a billion stars. By exploring the phase space (positions and velocities) of more than six million stars within a few kiloparsecs of the Sun in the Galactic disk from Gaia data release 2 (DR2; see Methods)¹⁹, we find that certain

phase-space projections (Figs. 1a, 2) have many substructures that had not been predicted by existing models. These substructures had remained blurred until now, owing to the limitations on the number of stars in and the precision of previously available datasets.

In Fig. 1a we show the projection of phase space in vertical position and velocity, Z – V_Z . The stars follow a curled, spiral-shaped distribution, the density of which increases towards the leading edge of the spiral. Figure 1b, c demonstrates that this ‘snail shell’ pattern is still present when the stars are colour-coded according to their radial and azimuthal velocities, V_R and V_ϕ , which implies a strong correlation between the vertical and in-plane motions of the stars. The pattern is particularly pronounced in the case of V_ϕ (Fig. 1c), even up to $V_Z \approx 40 \text{ km s}^{-1}$. Furthermore, we see a gradient of different values of V_ϕ across the spiral shape, which follows the density variations. Details of the relationship between the snail shell and the other velocity features observed in the solar neighbourhood are shown in Extended Data Fig. 1.

The spiral shape in Fig. 1a is clearly reminiscent of the effects of phase mixing in two dimensions that have been discussed in several areas of astrophysics^{20–22} and in quantum physics²³, but never in the context of dynamical models of the Galactic disk. This process can be better understood by using a toy model. Consider a Galaxy model in which the vertical potential of the Galaxy can be approximated by an anharmonic oscillator (see equation (1) in Methods). In this approximation, the vertical frequencies of oscillation ν depend on the amplitude of the oscillation A and the Galactocentric radius R , to first order²² (see equation (2) in Methods). By assuming that stars follow a simple harmonic oscillation with these frequencies, their movement over time t is described by $Z = A \cos[\phi(t)]$ and $V_Z = -A \nu \sin[\phi(t)]$ (where $\phi(t)$ is the orbital phase), which traces an oval shape in the clockwise direction in the Z – V_Z projection. However, stars revolve at different angular speeds depending on their frequency. Thus, an ensemble of stars will stretch out in phase space, with the range of frequencies causing a spiral

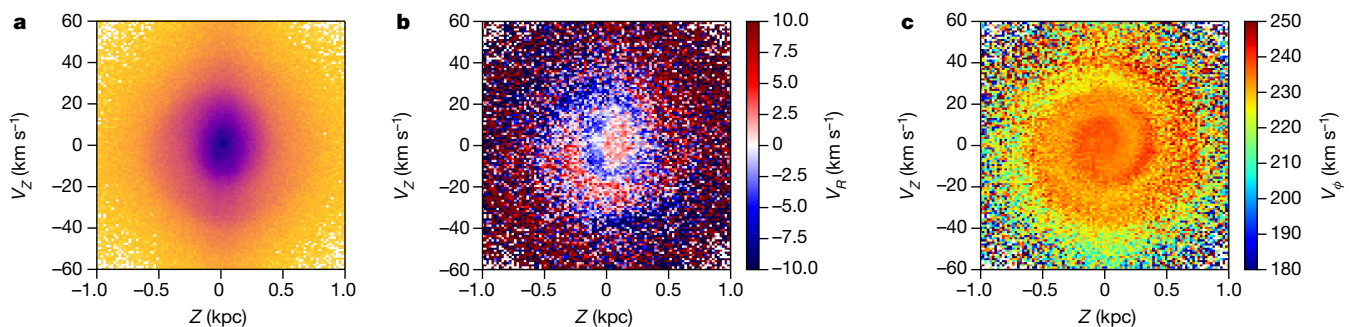


Fig. 1 | Vertical positions and velocities of the stars. The plots show the distribution of stars in the vertical position–velocity (Z – V_Z) plane from our sample of Gaia data for stars with Galactocentric radii of $8.24 \text{ kpc} < R < 8.44 \text{ kpc}$. **a**, Two-dimensional histogram in bins of $\Delta Z = 0.02 \text{ kpc}$ and $\Delta V_Z = 1 \text{ km s}^{-1}$, with the darkness of the colour scale

proportional to the number of stars. **b**, Z – V_Z plane coloured as a function of median radial velocity V_R in bins of $\Delta Z = 0.02 \text{ kpc}$ and $\Delta V_Z = 1 \text{ km s}^{-1}$. **c**, Same as **b**, but for the azimuthal velocity V_ϕ . V_R and V_ϕ are positive towards the Galactic anticentre and the direction of Galactic rotation, respectively.

¹Institut de Ciències del Cosmos, Universitat de Barcelona (IEEC-UB), Barcelona, Spain. ²Kapteyn Astronomical Institute, University of Groningen, Groningen, The Netherlands. ³GEPI, Observatoire de Paris, Université PSL, CNRS, Meudon, France. ⁴Université Grenoble Alpes, CNRS, IPAG, Grenoble, France. ⁵INAF—Osservatorio Astrofisico di Torino, Pino Torinese, Italy. ⁶Institute of Astronomy, University of Cambridge, Cambridge, UK. ⁷Università di Torino, Dipartimento di Fisica, Torino, Italy. ⁸Institut UTINAM, CNRS UMR6213, Université Bourgogne Franche-Comté, OSU THETA Franche-Comté Bourgogne, Observatoire de Besançon, Besançon, France. ⁹Mullard Space Science Laboratory, University College London, Dorking, UK. ¹⁰Laboratoire d’astrophysique de Bordeaux, Université Bordeaux, CNRS, Pessac, France. *e-mail: tantoja@fqa.ub.edu

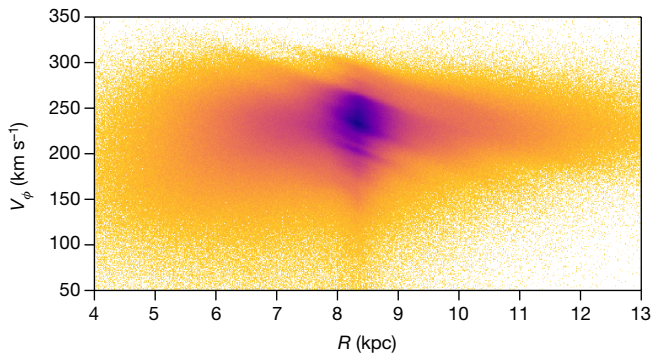


Fig. 2 | Positions and velocities of the stars in the disk plane.

Distribution of azimuthal velocities V_ϕ as a function of Galactocentric radius R for all stars in our sample with six-dimensional phase space-coordinates from Gaia DR2, shown as a two-dimensional histogram in bins of $\Delta V_\phi = 1 \text{ km s}^{-1}$ and $\Delta R = 0.01 \text{ kpc}$.

shape in this projection. The detailed time evolution of stars in this toy model is described in Methods and shown in Extended Data Fig. 3. As time goes by, the spiral gets more tightly wound and eventually this process of phase mixing leads to a spiral that is so tightly wound that the coarse-grained distribution appears to be smooth. The clarity of the spiral shape in the Z - V_Z plane revealed by Gaia DR2 implies that this time has not yet been reached in the Milky Way, providing evidence that phase mixing is currently taking place in the Galactic disk.

This interpretation also implies that the shape of the spiral can be used to obtain information about: (i) the shape of the potential, which determines the vertical frequencies; (ii) the starting time of phase mixing; and (iii) the type of perturbation that brought the disk into a non-equilibrium state, which sets the initial conditions for the phase-mixing event that we are witnessing. For instance, we can estimate the time t of the event from the separation between two consecutive spiral turns, because these have a phase separation of $(\nu_2 t + \phi_0) - (\nu_1 t + \phi_0) = 2\pi$, where the subscripts '1' and '2' indicate two consecutive turns of the spiral. Therefore, assuming that the initial phase ϕ_0 is the same for turns 1 and 2, we have $t = 2\pi/(\nu_2 - \nu_1)$. Using several potentials for the Milky Way (see Methods), we estimate that the vertical phase-mixing event started between 300 Myr and 900 Myr ago. The toy model that illustrates this process is shown in Fig. 3a, which depicts a snail shell that formed after 500 Myr from an ensemble of stars with a starting distribution that is out of equilibrium; this modelled snail shell is similar to the one seen in the data.

A possible perturbation that might have initiated the on-going vertical phase mixing that we observe is the influence of a satellite galaxy. In particular, the last pericentre of the orbit of the Sagittarius dwarf galaxy has been shown to have strong effects on the stellar disk^{4,8,9}. In addition, most models place this pericentric passage between 200 Myr and 1,000 Myr ago^{9,24,25}, which is consistent with our findings. Nevertheless, other processes that may induce snail shell patterns include the formation of the central bar and of the transient spiral structure, provided that these processes can induce vertical asymmetries, other global changes in the potential, and the dissolution of a massive stellar system such as a cluster or accreted satellite.

Another phase-space projection in which the Gaia data have a markedly different appearance is that in azimuthal velocity and cylindrical radius, V_ϕ - R (Fig. 2). Although this phase-space projection has been explored previously with other data²⁶, the spatial coverage, high sampling and unprecedented precision of the Gaia data reveal many thin diagonal ridges that were not previously evident. The arches in the velocity-space projection V_R - V_ϕ in the solar neighbourhood that were discovered recently in Gaia data¹² (see Extended Data Fig. 1a) are projections of these diagonal ridges, but at a fixed Galactic position. Therefore, Fig. 2 reveals that arched velocity structures must be present at many different radii, but have thus far been unexplored, and that their characteristics vary with distance from the Galactic centre,

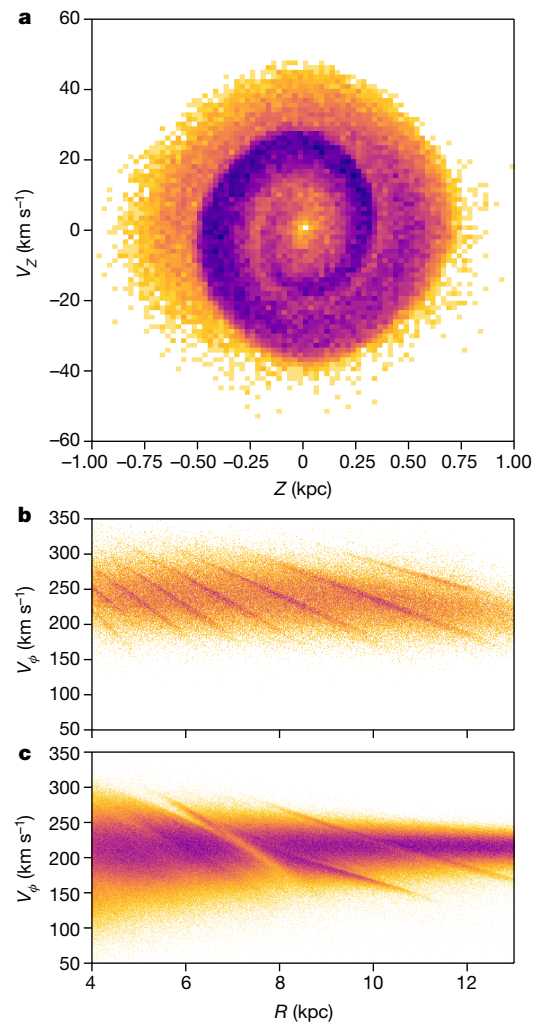


Fig. 3 | Models of the phase-space distribution of the Galaxy disk.

a, Modelled spiral shape created in the vertical position-velocity (Z - V_Z) plane as a result of phase mixing in the evolution of an ensemble of particles for 500 Myr in a Galactic potential, starting from a distribution that is out of equilibrium, presumably after a perturbation. **b**, Modelled diagonal ridges created in the distribution of azimuthal velocities V_ϕ as a function of Galactocentric radius R as a result of the phase mixing in the evolution of an ensemble of particles for 1,000 Myr in a Galactic potential, starting from a distribution that is out of equilibrium. **c**, Same as **b**, but for diagonal ridges created as a result of the effects of the barred potential and its resonant structure. See Methods for more details.

diminishing their velocity towards the outskirts of the Galaxy in a continuous way.

These diagonal ridges could be signatures of phase mixing in the horizontal direction, as has been predicted for the arches in velocity space⁶⁻⁸. Alternatively, the bar and the spiral arms could induce diagonal ridges through their resonant orbital structure, creating regions in phase space of stable and unstable orbits²⁷, and hence with over-densities and gaps. The toy model of phase mixing (Fig. 3b) and a disk simulation with a Galactic potential that contains a bar (Fig. 3c) both show several diagonal ridges. The V_ϕ separation of consecutive ridges in the data is about 10 km s^{-1} . Comparing this separation to that of our toy model (see Methods) indicates that, if these ridges are caused by phase mixing from a single perturbation, it should have taken place longer ago than the perturbation that gave rise to the vertical mixing. This is consistent with the timing derived using the separation between arches in the local velocity plane⁷ and with the existence of a group of co-moving stars that appear not to be fully phase mixed vertically¹⁸, which suggests that another perturbation occurred about 2 Gyr ago. The relationship between the various features is not clear, and it

is not unlikely that there are or were several perturbations creating superposed features.

Our interpretation of the features that we found is based on toy models, the main limitations of which are their lack of self-consistency, the choice of initial conditions not necessarily reflecting those stemming from the impact of a satellite galaxy, and the fact that we study separately both the effects of resonances and phase mixing, and the different phase-space dimensions (horizontal and vertical) involved. A challenging task for the future will be to model our findings taking into account collective effects, such as in the perturbative regime²⁸, and with self-consistent *N*-body models²⁵.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0510-7>.

Received: 26 April; Accepted: 26 July 2018;

Published online 19 September 2018.

- Sellwood, J. A. & Binney, J. J. Radial mixing in galactic discs. *Mon. Not. R. Astron. Soc.* **336**, 785–796 (2002).
- Contopoulos, G. & Grosbøl, P. Stellar dynamics of spiral galaxies: nonlinear effects at the 4/1 resonance. *Astron. Astrophys.* **155**, 11–23 (1986).
- Quinn, P. J., Hernquist, L. & Fullagar, D. P. Heating of galactic disks by mergers. *Astrophys. J.* **403**, 74–93 (1993).
- Purcell, C. W., Bullock, J. S., Tollerud, E. J., Rocha, M. & Chakrabarti, S. The Sagittarius impact as an architect of spirality and outer rings in the Milky Way. *Nature* **477**, 301–303 (2011).
- D’Onghia, E., Madau, P., Vera-Ciro, C., Quillen, A. & Hernquist, L. Excitation of coupled stellar motions in the Galactic disk by orbiting satellites. *Astrophys. J.* **823**, 4 (2016).
- Fux, R. Order and chaos in the local disc stellar kinematics induced by the Galactic bar. *Astron. Astrophys.* **373**, 511–535 (2001).
- Minchev, I. et al. Is the Milky Way ringing? The hunt for high-velocity streams. *Mon. Not. R. Astron. Soc.* **396**, L56–L60 (2009).
- Gómez, F. A., Minchev, I., Villalobos, A., O’Shea, B. W. & Williams, M. E. K. Signatures of minor mergers in Milky Way like disc kinematics: ringing revisited. *Mon. Not. R. Astron. Soc.* **419**, 2163–2172 (2012).
- de la Vega, A., Quillen, A. C., Carlin, J. L., Chakrabarti, S. & D’Onghia, E. Phase wrapping of epicyclic perturbations in the Wobbly galaxy. *Mon. Not. R. Astron. Soc.* **454**, 933–945 (2015).
- Eggen, O. J. Star streams and Galactic structure. *Astron. J.* **112**, 1595–1613 (1996).
- Dehnen, W. The distribution of nearby stars in velocity space inferred from HIPPARCOS data. *Astron. J.* **115**, 2384–2396 (1998).
- Gaia Collaboration. Gaia data release 2: mapping the Milky Way disc kinematics. *Astron. Astrophys.* **616**, A11 (2018).
- Siebert, A. et al. Detection of a radial velocity gradient in the extended local disc with RAVE. *Mon. Not. R. Astron. Soc.* **412**, 2026–2032 (2011).
- Widrow, L. M., Gardner, S., Yanny, B., Dodelson, S. & Chen, H.-Y. Galactoseismology: discovery of vertical waves in the Galactic disk. *Astrophys. J.* **750**, L41 (2012).
- Schönrich, R. & Dehnen, W. Warp, waves, and wrinkles in the Milky Way. *Mon. Not. R. Astron. Soc.* **478**, 3809–3824 (2018).
- Quillen, A. C. et al. The GALAH survey: stellar streams and how stellar velocity distributions vary with Galactic longitude, hemisphere and metallicity. *Mon. Not. R. Astron. Soc.* **478**, 228–254 (2018).
- Gómez, F. A. et al. Signatures of minor mergers in the Milky Way disc – I. The SEGUE stellar sample. *Mon. Not. R. Astron. Soc.* **423**, 3727–3739 (2012).
- Monari, G. et al. Cora Berenices: the first evidence for incomplete vertical phase-mixing in local velocity space with RAVE—confirmed with Gaia DR2. *Res. Notes AAS* **2**, 32 (2018).
- Gaia Collaboration. Gaia data release 2: summary of the contents and survey properties. *Astron. Astrophys.* **616**, A1 (2018).
- Tremaine, S. The geometry of phase mixing. *Mon. Not. R. Astron. Soc.* **307**, 877–883 (1999).
- Afshordi, N., Mohayaee, R. & Bertschinger, E. Hierarchical phase space structure of dark matter haloes: tidal debris, caustics, and dark matter annihilation. *Phys. Rev. D* **79**, 083526 (2009).
- Candlish, G. N. et al. Phase mixing due to the Galactic potential: steps in the position and velocity distributions of popped star clusters. *Mon. Not. R. Astron. Soc.* **437**, 3702–3717 (2014).
- Manfredi, G. & Feix, R. M. Theory and simulation of classical and quantum echoes. *Phys. Rev. E* **53**, 6460–6470 (1996).
- Law, D. R. & Majewski, S. R. The Sagittarius dwarf galaxy: a model for evolution in a triaxial Milky Way halo. *Astrophys. J.* **714**, 229–254 (2010).
- Laporte, C. F. P., Johnston, K. V., Gómez, F. A., Garavito-Camargo, N. & Besla, G. The influence of Sagittarius and the Large Magellanic Cloud on the Milky Way galaxy. *Mon. Not. R. Astron. Soc.* <https://doi.org/10.1093/mnras/sty1574> (2018).
- Monari, G., Kawata, D., Hunt, J. A. S. & Famaey, B. Tracing the Hercules stream with Gaia and LAMOST: new evidence for a fast bar in the Milky Way. *Mon. Not. R. Astron. Soc.* **466**, L113–L117 (2017).
- Michtchenko, T. A., Lépine, J. R. D., Barros, D. A. & Vieira, R. S. S. Combined dynamical effects of the bar and spiral arms in a Galaxy model. Application to the solar neighbourhood. *Astron. Astrophys.* **615**, A10 (2018).
- Fouvry, J.-B., Binney, J. & Pichon, C. Self-gravity, resonances, and orbital diffusion in stellar disks. *Astrophys. J.* **806**, 117 (2015).

Acknowledgements This work made use of data from ESA mission Gaia (<https://www.cosmos.esa.int/gaia>), which was processed by the Gaia Data Processing and Analysis Consortium (DPAC; <https://www.cosmos.esa.int/web/gaia/dpac/consortium>). Funding for the DPAC is provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement. This project received funding from the European Union’s Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement number 745617. This work was supported by the MDM-2014-0369 of ICCUB (Unidad de Excelencia ‘María de Maeztu’) and the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement GENIUS FP7-606740. A.H. acknowledges financial support from a VICI grant from the Netherlands Organisation for Scientific Research (NWO). We acknowledge the MINECO (Spanish Ministry of Economy) through grants ESP2016-80079-C2-1-R (MINECO/FEDER, UE) and ESP2014-55996-C2-1-R (MINECO/FEDER, UE). This work been funded in part by the Agenzia Spaziale Italiana (ASI) through contract 2014-025-R.1.2015 through the Italian Istituto Nazionale di Astrofisica (INAF). E.P. acknowledges the financial support of the 2014 PhD fellowship programme of INAF.

Author contributions T.A. contributed to the sample preparation, analysed and interpreted the data, performed most of the modelling and wrote the paper together with A.H. A.H. also provided interpretation of the findings. M.R.-G. performed the simulation with the barred potential and contributed to sample preparation. D.K., C.B., R.D., D.W.E., F.F., E.P., C.R., A.C.R., G.S. and C.S. contributed to sample preparation and validation of the Gaia data. All authors reviewed the manuscript.

Competing interests : The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0510-7>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to T.A.
Publisher’s note : Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data and samples selection. We used Gaia DR2 sources for which the six-dimensional phase-space coordinates can be computed, that is all stars for which a five-parameters astrometric solution (sky positions, parallax and proper motions) and a line-of-sight velocity are available. We selected only stars with positive parallaxes ϖ with relative uncertainty less than 20%, that is, satisfying $\varpi/\sigma_\varpi > 5$. This selection ensures that $1/\varpi$ is a reasonably good estimator of the distance to the stars²⁹; alternatively, we also used Bayesian distances (see below). This sample has 6,376,803 stars and has been well studied and characterized previously¹². The data were obtained directly through the following query in the public Gaia archive (<https://gea.esac.esa.int/archive>):

```
SELECT G.source_id, G.radial_velocity, G.radial_velocity_error, G.ra, G.ra_error,
G.dec, G.dec_error, G.parallax, G.parallax_error, G.pmra, G.pmra_error,
G.pmdec, G.pmdec_error, G.ra_dec_corr, G.ra_parallax_corr, G.ra_pmra_corr,
G.ra_pmdec_corr, G.dec_parallax_corr, G.dec_pmra_corr, G.dec_pmdec_corr,
G.parallax_pmra_corr, G.parallax_pmdec_corr, G.pmra_pmdec_corr
```

```
FROM gaiadr2.gai_source G
```

```
WHERE G.radial_velocity IS NOT Null AND G.parallax_over_error>5.
```

From the five-parameter astrometric solution and line-of-sight velocities ($\alpha, \delta, \varpi, \mu_\alpha^*, \mu_\delta, V_{\text{los}}$) of these stars, we derived distances (as $1/\varpi$), positions and velocities in the cylindrical Galactic reference frame ($R, \phi, Z, V_R, V_\phi, V_Z$). For convenience, we took ϕ to be positive in the direction of Galactic rotation and with the origin at the Sun–Galactic centre line. For these transformations, we adopted a vertical distance of the Sun above the plane of³⁰ 27 pc, a distance of the Sun to the Galactic centre³¹ of $R_\odot = 8.34$ kpc and a circular velocity at the Sun radius of³¹ $V_C(R_\odot) = 240$ km s⁻¹. We assumed a peculiar velocity of the Sun with respect to the local standard of rest of³² $(U_\odot, V_\odot, W_\odot) = (11.1, 12.24, 7.25)$ km s⁻¹. Our choice of values gives $(V_C(R_\odot) + V_\odot)/R_\odot = 30.2$ km s⁻¹ kpc⁻¹, which is compatible with the reflex motion of Sgr A³³. To derive the uncertainties in these coordinates, we propagate the full covariance matrix. The median uncertainties in V_R, V_ϕ and V_Z are 1.4 km s⁻¹, 1.5 km s⁻¹ and 1.0 km s⁻¹, respectively, and 80% of stars have uncertainty smaller than 3.3 km s⁻¹, 3.7 km s⁻¹ and 2.2 km s⁻¹, respectively, in these velocities. The positions in the Cartesian coordinates X – Y and X – Z of the sample are shown in Extended Data Fig. 2.

For part of our study, we selected from our sample the 935,590 stars located in the solar Galactic cylindrical ring, with Galactocentric radius 8.24 kpc $< R < 8.44$ kpc (dotted lines in Extended Data Fig. 2). For this selection, the median uncertainties in V_R, V_ϕ and V_Z are 0.5 km s⁻¹, 0.8 km s⁻¹ and 0.6 km s⁻¹, respectively, and 80% of stars have uncertainty smaller than 1.1 km s⁻¹, 2.0 km s⁻¹ and 1.3 km s⁻¹, respectively, in these velocities.

The velocity uncertainties are significantly smaller than the sizes of the substructures detected; together with the large number of stars in our samples, this is what made their detection possible. Although there are some correlations between the astrometric Gaia observables³⁴, these are not responsible for the correlations and substructures seen in our phase-space plots. This is because the stars in our sample are distributed across all sky directions, and the phase-space coordinates come from combinations of astrometric measurements and radial velocities in different contributions depending on the direction on the sky. Besides, the astrometric correlations for our sample are small (less than 0.2 in their absolute value for more than 50% of stars); this, combined with the small errors, makes their contribution insignificant.

Alternatively, we used distances determined through a Bayesian inference method using the existing implementation in TOPCAT³⁵, taking the mode of the posterior distribution and a prior of an exponentially decreasing density of stars with scale length of³⁶ 1.35 kpc. We found that the differences between this distance determination and the inverse of the parallax are between -2% and 0.6% for 90% of the 6,376,803 stars with $\varpi/\sigma_\varpi > 5$, which is expected for small relative errors in parallax. Consequently, the phase-space diagrams presented here vary only at the pixel level. The diagrams do not vary even when using the set of 7,183,262 of stars with available radial velocities, which includes stars with larger parallax errors and with negative parallaxes, for which the estimator of the inverse of the parallax would yield unphysical distances. When using another alternative set of Bayesian distances specifically derived for stars from Gaia DR2 with radial velocities using a different prior³⁷, we found the differences between these distances and the inverse of the parallax to be between -9% and 5% for 90% of the stars—slightly larger than before, but again with no noticeable effects on the phase-space projections examined here.

Models for vertical phase mixing. We first reproduced the spiral shape observed in the Z – V_Z plane with Gaia data by using a toy model. The classic harmonic oscillator is often used to describe the vertical movement of stars in galactic disks under the epicyclic theory³⁸. However, in this approximation, which is valid only

for very small-amplitude orbits for which the potential changes little vertically, stars have the same vertical oscillatory frequency ν and there is no phase mixing, unless orbits at different guiding radius, and therefore with different frequencies, are considered. Instead, we use an anharmonic oscillator with potential

$$\Phi(Z) \propto -\alpha_0 + \frac{1}{2}\alpha_1 Z^2 - \frac{1}{4}\alpha_2 Z^4 \quad (1)$$

We adopt coefficients α_0, α_1 and α_2 that correspond to the expansion for small Z (derived elsewhere²²) of a Miyamoto–Nagai potential³⁹ with $a = 6.5$ kpc, $b = 0.26$ kpc and $M = 10^{11} M_\odot$. These coefficients depend on Galactocentric radius R because the vertical pull depends on the distance to the Galactic centre. In this anharmonic potential, the frequencies of oscillation are described by

$$\nu(A, R) = \alpha_1(R)^{1/2} \left[1 - \frac{3\alpha_2(R)A^2}{8\alpha_1(R)} \right] \quad (2)$$

where $\nu_0 = \alpha_1^{1/2}$ is the vertical frequency in the epicyclic approximation.

Given an initial distribution of stars with $Z(t=0)$ and $V_Z(t=0)$, the vertical amplitudes of the orbits can be derived through the conservation of energy and using the fact that at the vertical turn-around point of the orbit ($V_Z=0$) the (vertical) kinetic energy is null²². Assuming that stars follow a simple harmonic oscillation (but with different frequencies), the movement of the stars with time is described by

$$\begin{aligned} Z &= A \cos[\nu(A, R)t + \phi_0] \\ V_Z &= -A\nu(A, R)\sin[\nu(A, R)t + \phi_0] \end{aligned} \quad (3)$$

where the initial phase of the stars $\phi_0 = \phi(t=0)$ is obtained from the initial distribution of Z and V_Z and the corresponding amplitudes.

The phase-space evolution described above is shown in Extended Data Fig. 3a–c. Initially, the particles follow a Gaussian distribution in $Z(t=0)$ and $V_Z(t=0)$ with mean and dispersion of -0.1 kpc and 0.04 kpc, and -2 km s⁻¹ and 1 km s⁻¹, respectively. We located all particles at the same Galactocentric radius of $R = 8.5$ kpc; therefore, they all move under the same functional form of the vertical potential. The initial conditions are shown in Extended Data Fig. 3a, in which we have colour-coded the particles according to their period. Following equation (3), each star follows a clockwise rotation in the Z – V_Z plane. However, they do so at different angular speeds: stars with smaller period located at closer distances from the mid-plane ($Z=0$) revolve faster than those located at largest distances from the mid-plane. The range of frequencies is therefore what creates the spiral shape. Extended Data Fig. 3b shows the evolution of the system for three initial phases of the time evolution when the spiral shape begins to form. Extended Data Fig. 3c shows the spiral shape after 1,000 Myr of evolution.

In the Gaia data (Fig. 1), we do not see a thin spiral but a thick one, with many of the stars in the volume participating in it. A similar effect was observed with our toy model when we included particles at different radius for which the vertical potential and the range of amplitudes and frequencies changes. In Extended Data Fig. 3d–f we let a similar system evolve as in Extended Data Fig. 3a–c but starting with the initial radius following a skewed normal distribution, which creates a density that decreases with radius, as in galaxy disks, with a skewness of 10, location parameter of 8.4 kpc and scale parameter of 0.2 kpc. The spiral structure is now thickened, similarly to the data, with a higher density of stars at the leading edge of the spiral.

To estimate the time of the phase-mixing event from the spiral seen in the Gaia data (Fig. 1) using

$$t = \frac{2\pi}{\nu_2 - \nu_1} \quad (4)$$

we need to locate two consecutive turns of the spiral and estimate their vertical frequencies from their amplitudes and mean radius. For this, we used Extended Data Fig. 5, which is colour-coded as a function of median guiding radius. The guiding radius is approximated as $R_g \approx V_\phi R_\odot / [V_C(R_\odot)]$, under the hypothesis of a flat rotation curve, where we used the values of $R_\odot = 8.34$ kpc and $V_C(R_\odot) = 240$ km s⁻¹ assumed in the coordinate transformation of the data. From Extended Data Fig. 5 we see that the density gradient across the spiral shape is created by stars with different guiding radius that arrive at the solar neighbourhood owing to their different amplitudes of (horizontal) radial oscillation. To determine two consecutive turns of the spiral, we focused on stars at the turn-around points ($V_Z=0$) near the leading edges of the spiral. By visual inspection, we determined an approximate range of Z in which the turn-around points are located in Extended Data Fig. 5, concentrating on red colours, for which the spiral is well defined. The ranges of the turn-around points are marked with vertical lines and listed, together with the middle value, in Extended Data Table 1. For these turn-around points, the

amplitudes are simply $A = Z$ and from the colour bar we note that the median R_g is around 8.2 kpc. Small changes in this value do not change substantially our final determination of the time of the perturbation.

To estimate the vertical orbital frequencies of these turn-around points, we could not use the toy model presented above because it is valid only for oscillations with small amplitude A —in particular, smaller than the vertical scale b of the potential ($A \ll 0.26$ kpc). Therefore, we took an existing model⁴⁰ with updated parameters that fit current estimates such as for the Sun Galactocentric radius and the circular velocity curve⁴¹. We computed the vertical frequency numerically in a grid of different radii and vertical amplitudes by integrating orbits and measuring their vertical periods (Extended Data Fig. 4a). The vertical frequency can change along the orbits for stars with large eccentricities in the horizontal direction, but here for simplicity we put all particles on nearly circular orbits. We estimated the vertical frequency at each turn-around position directly by interpolating the calculations of Extended Data Fig. 4a using the estimated values for the amplitude and radius.

Finally, taking each pair of turning points, we obtained an estimate of the time since the perturbation using equation (4). As an example, the two turning points (amplitudes) of the left part of the spiral are located at -0.59 ± 5 kpc and -0.23 ± 5 kpc, respectively. These correspond to vertical frequencies of 0.058 ± 0.002 rad Myr⁻¹ and 0.072 ± 0.002 rad Myr⁻¹ for $R_g = 8.2$ kpc, which gives a time of 461^{+183}_{-105} Myr. We repeated the same procedure for the second pair of consecutive turning points, and also for the edges of the spiral at $Z = 0$ (mid-plane points, which have $V_z = A\nu$), estimating the frequencies by interpolating in Extended Data Fig. 4b. The mid-plane positions are marked as horizontal lines in Extended Data Fig. 5. All results are summarized in Extended Data Tables 1 and 2. The mean of the three time estimates is 510 Myr and the minimum and maximum times from the uncertainty ranges are 356 Myr and 856 Myr.

We tested the dependence of our time determination on the potential model used by using a different model⁴². Compared to our previous model, this one has different shapes for the halo, disks and bulge, a different total mass, and includes thin and thick disks and two gas disks. The frequencies in this model are smaller by 4% on average and smaller than 6% for 90% of the points in the grid of Extended Data Fig. 4a. By repeating the whole process to determine the perturbation time, we obtained 528 Myr with minimum and maximum times from the uncertainty ranges of 361 Myr and 899 Myr, very similar to our previous determination.

Our determination is subject to several approximations: (1) we used the vertical frequencies of orbits with conditions of circularity on the plane and for certain assumed Galactic potential models; (2) we considered a unique guiding radius; and (3) we took equal initial phases for the turn-around and mid-plane points.

We finally ran a simulation (Fig. 3a) by integrating 100,000 test-particle orbits in the updated model⁴¹ with initial vertical positions and velocities following Gaussian distributions centred at $Z = -0.4$ kpc and $V_z = -5$ km s⁻¹ and with dispersions of 0.15 kpc and 2 km s⁻¹, respectively. Horizontally in the disk plane, the test particles were distributed following a skewed normal distribution in radius R , with a scale parameter of 0.8 kpc, location parameter of 8, skewness of 10, and all particles at an azimuthal angle $\phi = 0$. The horizontal velocities were set to 0 for the radial component and to the circular velocity at the particle's radius and $Z = 0$ for the azimuthal component. These are initial conditions of circularity on the Galactic plane but not necessarily for orbits with large excursions in Z . The particle's orbits were integrated forwards in time for 500 Myr as estimated from the data. This is not meant to be a fit to the data because we have not explored all possible initial configurations that could lead to a similar spiral shape. However, we see that an initial distribution that is asymmetric in Z , with most particles located at positive or negative Z , is required to obtain a single spiral instead of a symmetrical double one.

Model for the horizontal phase mixing. We used a toy model to reproduce the diagonal ridges observed in the V_ϕ - R plane (Fig. 3b). This model is built by integrating orbits in the Galactic potential of ref.⁴⁰ with updated parameters that fit current estimates such as for the Sun Galactocentric radius or the circular velocity curve⁴¹. We used as initial conditions a set of test particles distributed in Galactocentric radius according to a skewed normal distribution with a skewness of 10, location parameter of 4 kpc and scale parameter of 6 kpc. The azimuthal angle was fixed between 0° and 50° to mimic a localized perturbation in the disk. For simplicity, all particles were put at the mid-plane with null vertical velocities. The radial and azimuthal velocities were initialized, respectively, following Gaussian distributions centred at 0 with a dispersion of 40 km s⁻¹ and centred at the circular velocity at the particular radius with a dispersion of 30 km s⁻¹. The particle orbits were computed for 1 Gyr. To these particles we added particles in the disk that had been integrated for a much longer timescale and are therefore well mixed, to

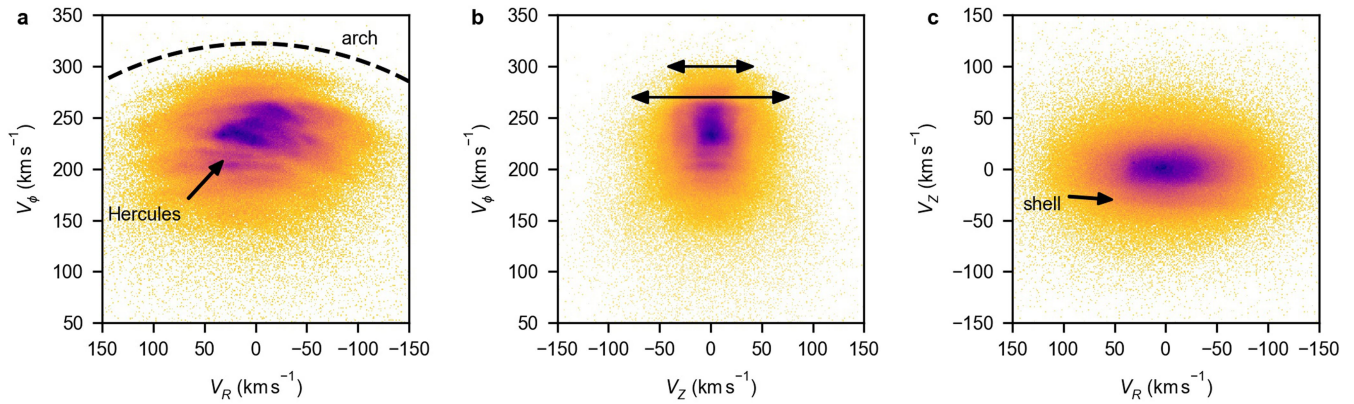
simulate the stellar populations that have not been perturbed. Of all the particles in the simulation, we used only the ones located in a range of 10° in azimuthal angle at the end of the integration, similarly to the data.

Model for the horizontal resonances. The model of Fig. 3c is from a test-particle simulation of orbits integrated in a Galactic potential model including a bar⁴³. The axisymmetric part of the potential was from an existing model⁴⁰. The Galactic bar potential was built using Ferrers ellipsoids⁴⁴ oriented with semi-major axes at 20° from the Sun–Galactic centre line, and with pattern speed set to 50 km s⁻¹ kpc⁻¹, which corresponds to a period of about 120 Myr. The simulation consisted of 68 million test particles with an initial radial velocity dispersion of 30 km s⁻¹ at the solar radius. Their orbits were first integrated in the axisymmetric potential model for 10 Gyr until they were approximately fully phase-mixed. Next, the bar potential was grown in $T_{\text{grow}} = 4$ rotations of the bar. More details on the bar potential, initial conditions and integration procedure are specified elsewhere⁴³. Here we used the final conditions after T_{grow} (about 500 Myr) and eight additional bar rotations (about 1,000 Myr). From all of the particles in the simulation, we used only the ones located in a range of 10° in azimuthal angle centred on the Sun, similarly to the data.

Code availability. We have made use of standard data analysis tools in Python. The codes used to generate the toy models and simulations and to compute the orbital frequencies are available from the corresponding author on reasonable request. The code used to compute the orbits for the potential from ref.⁴² is available at <https://github.com/PaulMcMillan-Astro/GalPot>. The code used to compute Bayesian distances from parallaxes is available in the TOPCAT platform³⁵.

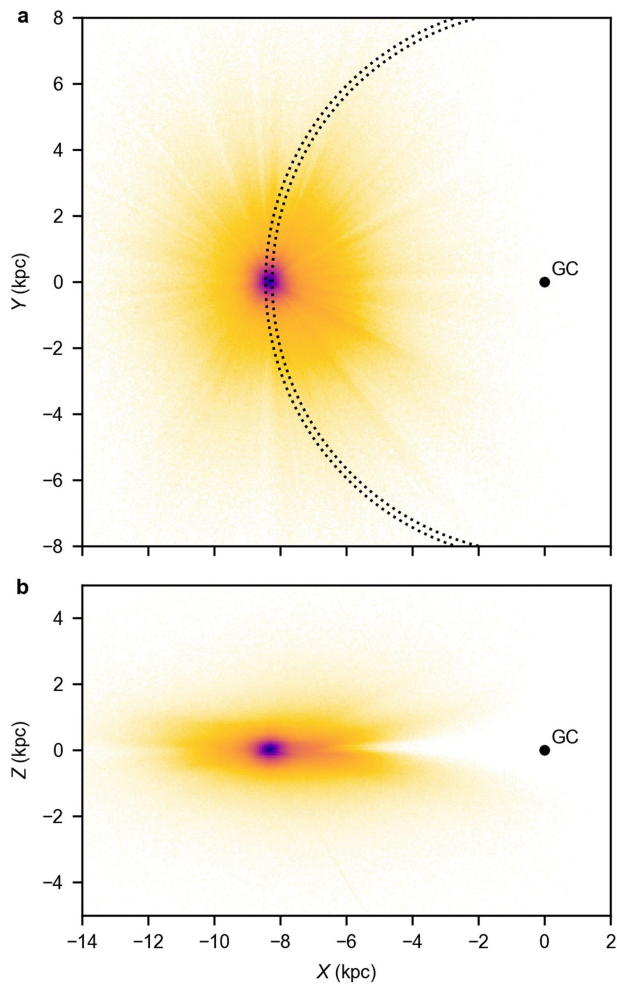
Data availability. The datasets used and analysed for this study are derived from data available in the public Gaia archive (<https://gea.esac.esa.int/archive/>). The Bayesian distances for the Gaia sources with radial velocity³⁷ are available at http://www.astro.lu.se/~paul/GaiaDR2_RV_star_distance.csv.gz. The rest of the relevant datasets and toy models are available from the corresponding author on reasonable request.

29. Luri, X. et al. Gaia data release 2: using Gaia parallaxes. *Astron. Astrophys.* **616**, A9 (2018).
30. Chen, B. et al. Stellar population studies with the SDSS. I. The vertical distribution of stars in the Milky Way. *Astrophys. J.* **553**, 184–197 (2001).
31. Reid, M. J. et al. Trigonometric parallaxes of high mass star forming regions: the structure and kinematics of the Milky Way. *Astrophys. J.* **783**, 130 (2014).
32. Schönrich, R. Galactic rotation and solar motion from stellar kinematics. *Mon. Not. R. Astron. Soc.* **427**, 274–287 (2012).
33. Reid, M. J. & Brunthaler, A. The proper motion of Sagittarius A*. II. The mass of Sagittarius A*. *Astrophys. J.* **616**, 872–884 (2004).
34. Lindegren, L. et al. Gaia data release 2: the astrometric solution. *Astron. Astrophys.* **616**, A2 (2018).
35. Taylor, M. B. TOPCAT & STIL: Starlink table/VOTable processing software. *ASP Conf. Ser.* **347**, 29–33 (2005).
36. Astraatmadja, T. L. & Bailer-Jones, C. A. L. Estimating distances from parallaxes. II. Performance of Bayesian distance estimators on a Gaia-like catalogue. *Astrophys. J.* **832**, 137 (2016).
37. McMillan, P. J. Simple distance estimates for Gaia DR2 stars with radial velocities. *Res. Notes AAS* **2**, 51 (2018).
38. Binney, J. & Tremaine, S. *Galactic Dynamics* 2nd edn (Princeton Univ. Press, Princeton, 2008).
39. Miyamoto, M. & Nagai, R. Three-dimensional models for the distribution of mass in galaxies. *Publ. Astron. Soc. Jpn* **27**, 533–543 (1975).
40. Allen, C. & Santillan, A. An improved model of the galactic mass distribution for orbit computations. *Rev. Mex. Astron. Astrofis.* **22**, 255–263 (1991).
41. Irigang, A., Wilcox, B., Tucker, E. & Schiefelbein, L. Milky Way mass models for orbit calculations. *Astron. Astrophys.* **549**, A137 (2013).
42. McMillan, P. J. The mass distribution and gravitational potential of the Milky Way. *Mon. Not. R. Astron. Soc.* **465**, 76–94 (2017).
43. Romero-Gómez, M., Figueras, F., Antoja, T., Abedi, H. & Aguilar, L. The analysis of realistic stellar Gaia mock catalogues – I. Red clump stars as tracers of the central bar. *Mon. Not. R. Astron. Soc.* **447**, 218–233 (2015).
44. Ferrers, N. On the potentials of ellipsoids, ellipsoidal shells, elliptic laminae and elliptic rings of variable densities. *QJ Pure Appl. Math.* **14**, 1–22 (1877).
45. Eggen, O. J. Stellar groups. II. The ζ Herculis, ϵ Indi and 61 Cygni groups of high-velocity stars. *Mon. Not. R. Astron. Soc.* **118**, 154–160 (1958).
46. Blaauw, A. Remarks on Local Structure and Kinematics. *Symp. IAU* **38**, 199–204 (1970).
47. Skuljan, J., Hearnshaw, J. B. & Cottrell, P. L. Velocity distribution of stars in the solar neighbourhood. *Mon. Not. R. Astron. Soc.* **308**, 731–740 (1999).
48. Antoja, T., Figueras, F., Fernández, D. & Torra, J. Origin and evolution of moving groups. I. Characterization in the observational kinematic-age-metallicity space. *Astron. Astrophys.* **490**, 135–150 (2008).

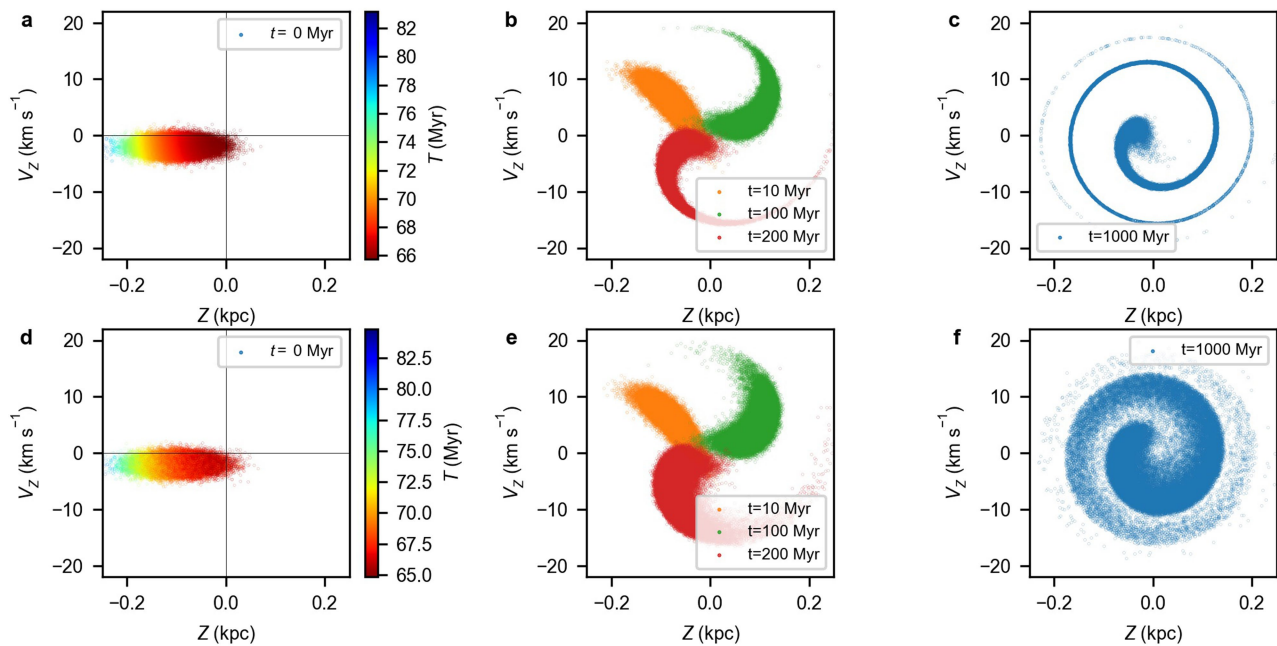


Extended Data Fig. 1 | Velocities of the stars at the solar Galactocentric radius. Two-dimensional histograms of combinations of radial, azimuthal and vertical Galactic cylindrical velocities for the stars in our sample of Gaia data located at $8.24 \text{ kpc} < R < 8.44 \text{ kpc}$, in bins of 1 km s^{-1} . V_R and V_ϕ are positive towards the Galactic anticentre and the direction of Galactic rotation, respectively. The darkness is proportional to the number of counts. **a**, Although the bimodality seen here, separating the Hercules stream from the rest of the distribution, was known^{45,46}, as well as some other elongated structures in this velocity projection^{11,47,48}, the numerous and thin arches are a new phenomenon revealed by Gaia data¹². The

semi-circular dotted line marks an arbitrary line of constant kinetic energy in the plane $E_k = (V_R^2 + V_\phi^2)/2$, as predicted for the substructure generated in horizontal phase mixing^{7,8}. **b**, The data have a box-like appearance, where the extent in V_z of the arches varies with V_ϕ (arrows), probably created by the correlation between the spiral shape and V_ϕ seen in Fig. 1c. **c**, Although some velocity asymmetries were noticed before in the V_ϕ - V_z projection¹¹ and attributed to the Galaxy warp, the sharp shell-like features involving V_z , especially at $V_z \approx -30 \text{ km s}^{-1}$ and $V_z \approx 25 \text{ km s}^{-1}$, were not previously evident. These shells are different projections of the snail shell pattern of Fig. 1a.

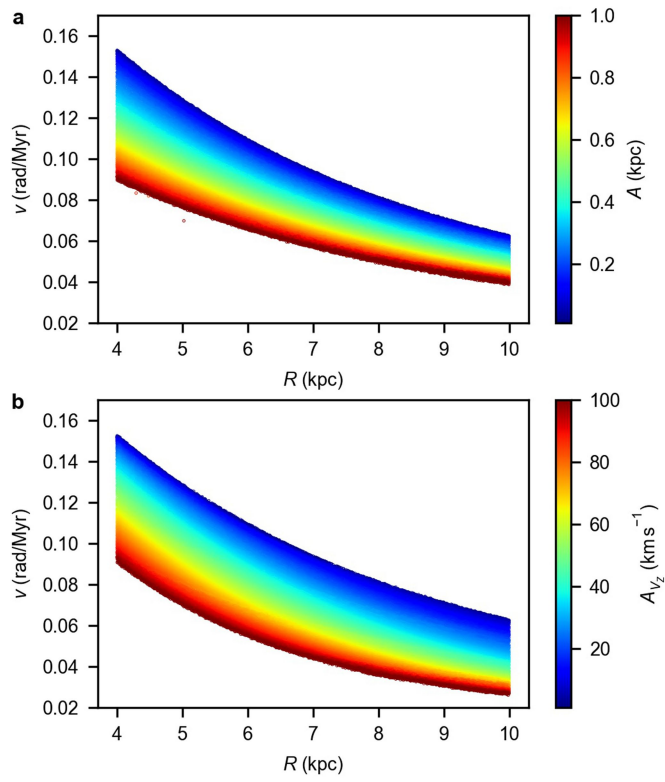


Extended Data Fig. 2 | Location of the stars in the sample. **a, b,** Two-dimensional histograms with bins of 0.05 kpc in the X - Y (**a**) and X - Z (**b**) projections of our sample of Gaia data. The dotted lines mark the selection of stars in the solar Galactic ring between radii of 8.24 kpc and 8.44 kpc. The Sun is located at $(X, Y, Z) = (-8.34, 0, 0.027)$ kpc and the Galactic centre (GC) is marked with a black dot.

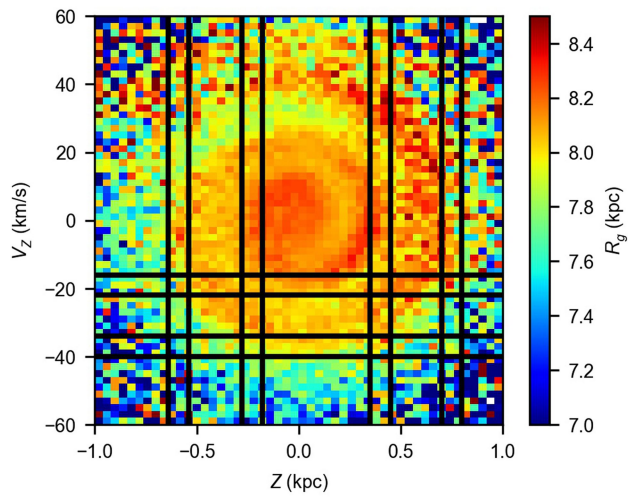


Extended Data Fig. 3 | Modelled vertical positions and velocities of stars with time. The plots show the snail shells created in the phase space evolution under an anharmonic potential. **a–c**, Phase-space evolution at different times ($t = 0, 10, 100, 200, 1,000$ Myr) for an ensemble of particles at a fixed Galactocentric radius of $R = 8.5$ kpc with an initial Gaussian distribution in $Z(t=0)$ with mean of -0.1 kpc and dispersion of 0.04 kpc

and in $V_Z(t=0)$ with mean of -2 km s^{-1} and dispersion of 1 km s^{-1} . **d–f**, Same as **a–c**, but for a skewed normal distribution of initial radius with skewness of 10, location parameter of 8.4 kpc and scale parameter of 0.2 kpc. In all cases, the evolution is under an anharmonic oscillator derived from the expansion of a Miyamoto–Nagai disk for small Z . In **a** and **d** the stars are colour-coded by vertical period.



Extended Data Fig. 4 | Vertical frequency for orbits in a Galaxy model.
a, b, Frequencies as a function of Galactocentric radius R computed in the updated model from ref. ⁴¹, colour coded by the vertical amplitude (**a**) and by the vertical velocity amplitude (**b**) of the orbits.



Extended Data Fig. 5 | Position of the spiral turns in the vertical positions and velocities. The Z - V_Z plane for stars at Galactocentric radii of 8.24 kpc to 8.44 kpc, coloured as a function of median guiding radius R_g in bins of $\Delta Z = 0.04$ kpc and $\Delta V_Z = 2$ km s $^{-1}$, with vertical and horizontal lines showing the approximate locations of the observed snail shell (turn-around and mid-plane points).

Extended Data Table 1 | Time estimates from the turn-around points of the spiral

Z (kpc)		ν (rad/Myr)	Time (Myr)
-0.59	± 5	0.058 ± 0.002	461^{+183}_{-105}
-0.23	± 5	0.072 ± 0.002	
0.40	± 5	0.065 ± 0.003	566^{+290}_{-140}
0.75	± 5	0.054 ± 0.001	

The first column indicates the vertical positions of the turn-around points, which are equal to the amplitude of the orbits except for the sign, and the estimated uncertainty ranges. The other columns are the frequencies corresponding to these amplitudes and the starting times of the phase-mixing process corresponding to each pair of consecutive spiral turns.

Extended Data Table 2 | Time estimates from the mid-plane points of the spiral

A_{v_z} (km s^{-1})		ν (rad/Myr)	Time (Myr)
-37	± 3	0.066 ± 0.002	505^{+253}_{-132}
-19	± 3	0.079 ± 0.002	

The first column indicates the vertical velocities at the mid-plane passages, which are equal to the velocity amplitudes of the orbits except for the sign, and the estimated uncertainty ranges. The other columns are the frequencies corresponding to these amplitudes and the starting time of the phase-mixing process corresponding to the pair of consecutive spiral turns.

Acceleration of electrons in the plasma wakefield of a proton bunch

E. Adli¹, A. Ahuja², O. Apsimon^{3,4}, R. Apsimon^{4,5}, A.-M. Bachmann^{2,6,7}, D. Barrientos², F. Batsch^{2,6,7}, J. Bauche², V. K. Berglyd Olsen¹, M. Bernardini², T. Bohl², C. Bracco², F. Braummüller⁶, G. Burt^{4,5}, B. Buttenschön⁸, A. Caldwell⁶, M. Cascella⁹, J. Chappell⁹, E. Chevallay², M. Chung¹⁰, D. Cooke⁹, H. Damerau², L. Deacon⁹, L. H. Deubner¹¹, A. Dexter^{4,5}, S. Doebert², J. Farmer¹², V. N. Fedosseev², R. Fiorito^{4,13}, R. A. Fonseca¹⁴, F. Friebe², L. Garolfi², S. Gessner², I. Gorgisyan², A. A. Gorn^{15,16}, E. Granados², O. Grulke^{8,17}, E. Gschwendtner², J. Hansen², A. Helm¹⁸, J. R. Henderson^{4,5}, M. Hüther⁶, M. Ibison^{4,13}, L. Jensen², S. Jolly⁹, F. Keeble⁹, S.-Y. Kim¹⁰, F. Kraus¹¹, Y. Li^{3,4}, S. Liu¹⁹, N. Lopes¹⁸, K. V. Lotov^{15,16}, L. Maricalva Brun², M. Martyanov⁶, S. Mazzoni², D. Medina Godoy², V. A. Minakov^{15,16}, J. Mitchell^{4,5}, J. C. Molendijk², J. T. Moody⁶, M. Moreira^{2,18}, P. Muggli^{2,6}, E. Öz⁶, C. Pasquino², A. Pardons², F. Peña Asmus^{6,7}, K. Pepitone², A. Perera^{4,13}, A. Petrenko^{2,15}, S. Pitman^{4,5}, A. Pukhov¹², S. Rey², K. Rieger⁶, H. Ruhl²⁰, J. S. Schmidt², I. A. Shalimova^{16,21}, P. Sherwood⁹, L. O. Silva¹⁸, L. Soby², A. P. Sosedkin^{15,16}, R. Speroni², R. I. Spitsyn^{15,16}, P. V. Tuev^{15,16}, M. Turner², F. Velotti², L. Verra^{2,22}, V. A. Verzilov¹⁹, J. Vieira¹⁸, C. P. Welsch^{4,13}, B. Williamson^{3,4}, M. Wing^{9*}, B. Woolley² & G. Xia^{3,4}

High-energy particle accelerators have been crucial in providing a deeper understanding of fundamental particles and the forces that govern their interactions. To increase the energy of the particles or to reduce the size of the accelerator, new acceleration schemes need to be developed. Plasma wakefield acceleration^{1–5}, in which the electrons in a plasma are excited, leading to strong electric fields (so called ‘wakefields’), is one such promising acceleration technique. Experiments have shown that an intense laser pulse^{6–9} or electron bunch^{10,11} traversing a plasma can drive electric fields of tens of gigavolts per metre and above—well beyond those achieved in conventional radio-frequency accelerators (about 0.1 gigavolt per metre). However, the low stored energy of laser pulses and electron bunches means that multiple acceleration stages are needed to reach very high particle energies^{5,12}. The use of proton bunches is compelling because they have the potential to drive wakefields and to accelerate electrons to high energy in a single acceleration stage¹³. Long, thin proton bunches can be used because they undergo a process called self-modulation^{14–16}, a particle–plasma interaction that splits the bunch longitudinally into a series of high-density microbunches, which then act resonantly to create large wakefields. The Advanced Wakefield (AWAKE) experiment at CERN^{17–19} uses high-intensity proton bunches—in which each proton has an energy of 400 gigaelectronvolts, resulting in a total bunch energy of 19 kilojoules—to drive a wakefield in a ten-metre-long plasma. Electron bunches are then injected into this wakefield. Here we present measurements of electrons accelerated up to two gigaelectronvolts at the AWAKE experiment, in a demonstration of proton-driven plasma wakefield acceleration. Measurements were conducted under various plasma conditions and the acceleration was found to be consistent and reliable. The potential for this scheme to produce very high-energy electron bunches in a single accelerating stage²⁰ means that our results are an important step towards the development of future high-energy particle accelerators^{21,22}.

The layout of the AWAKE experiment is shown in Fig. 1. A proton bunch from CERN’s Super Proton Synchrotron (SPS) accelerator co-propagates with a laser pulse (green), which creates a plasma (yellow) in a column of rubidium vapour (pink) and seeds the

modulation of the proton bunch into microbunches (Fig. 1; red, bottom images). The protons have an energy of 400 GeV and the root-mean-square (r.m.s.) bunch length is 6–8 cm¹⁸. The bunch is focused to a transverse size of approximately 200 μm (r.m.s.) at the entrance of the vapour source, with the bunch population varying shot-to-shot in the range $N_p \approx (2.5–3.1) \times 10^{11}$ protons per bunch. Proton extraction occurs every 15–30 s. The laser pulse used to singly ionize the rubidium in the vapour source^{23,24} is 120 fs long with a central wavelength of 780 nm and a maximum energy of 450 mJ²⁵. The pulse is focused to a waist of approximately 1 mm (full-width at half-maximum, FWHM) inside the rubidium vapour source, five times the transverse size of the proton bunch. The rubidium vapour source (Fig. 1; centre) has a length of 10 m and diameter of 4 cm, with rubidium flasks at each end. The rubidium vapour density and hence the plasma density n_{pe} can be varied in the range $10^{14}–10^{15}$ cm^{−3} by heating the rubidium flasks to temperatures of 160–210 °C. This density range corresponds to a plasma wavelength of 1.1–3.3 mm, as detailed in Methods. A gradient in the plasma density can be introduced by heating the rubidium flasks to different temperatures. Heating the downstream (Fig. 1; right side) flask to a higher temperature than the upstream (left side) flask creates a positive density gradient, and vice versa. Gradients in plasma density have been shown in simulation to produce large increases in the maximum energy attainable by the injected electrons²⁶. The effect of density gradients here is different from that for short drivers²⁷. In addition to keeping the wake travelling at the speed of light at the witness position, the gradient prevents destruction of the bunches at the final stage of self-modulation²⁸, thus increasing the wakefield amplitude at the downstream part of the plasma cell. The rubidium vapour density is monitored constantly by an interferometer-based diagnostic²⁹.

The self-modulation of the proton bunch into microbunches (Fig. 1; red, bottom right image) is measured using optical and coherent transition radiation diagnostics (Fig. 1; purple)³⁰. However, these diagnostics have a destructive effect on the accelerated electron bunch and cannot be used during electron acceleration experiments. The second beam-imaging station (Fig. 1; orange, right) is used instead, providing an indirect measurement of the self-modulation by measuring the transversely defocused protons³¹. These protons are expelled from the

¹University of Oslo, Oslo, Norway. ²CERN, Geneva, Switzerland. ³University of Manchester, Manchester, UK. ⁴Cockcroft Institute, Daresbury, UK. ⁵Lancaster University, Lancaster, UK. ⁶Max Planck Institute for Physics, Munich, Germany. ⁷Technical University Munich, Munich, Germany. ⁸Max Planck Institute for Plasma Physics, Greifswald, Germany. ⁹UCL, London, UK. ¹⁰UNIST, Ulsan, South Korea. ¹¹Philipps-Universität Marburg, Marburg, Germany. ¹²Heinrich-Heine-University of Düsseldorf, Düsseldorf, Germany. ¹³University of Liverpool, Liverpool, UK. ¹⁴ISCTE—Instituto Universitário de Lisboa, Lisbon, Portugal. ¹⁵Budker Institute of Nuclear Physics SB RAS, Novosibirsk, Russia. ¹⁶Novosibirsk State University, Novosibirsk, Russia. ¹⁷Technical University of Denmark, Lyngby, Denmark. ¹⁸GoLP/Instituto de Plasmas e Fusão Nuclear, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal. ¹⁹TRIUMF, Vancouver, British Columbia, Canada. ²⁰Ludwig-Maximilians-Universität, Munich, Germany. ²¹Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russia. ²²University of Milan, Milan, Italy.

*e-mail: m.wing@ucl.ac.uk

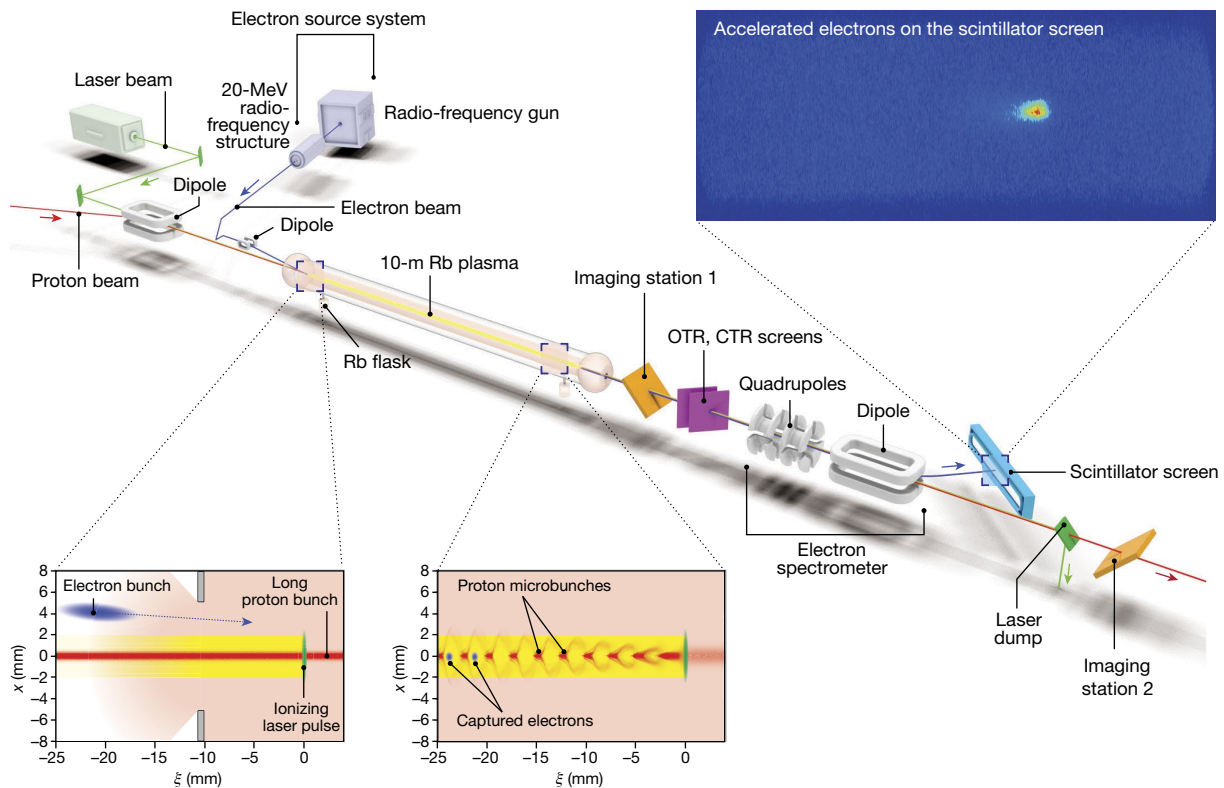


Fig. 1 | Layout of the AWAKE experiment. The proton bunch and laser pulse propagate from left to right across the image, through a 10-m column of rubidium (Rb) vapour. This laser pulse (green, bottom images) singly ionizes the rubidium to form a plasma (yellow), which then interacts with the proton bunch (red, bottom left image). This interaction modulates the long proton bunch into a series of microbunches (bottom right image), which drive a strong wakefield in the plasma. These microbunches are millimetre-scale in the longitudinal direction (ξ) and submillimetre-scale in the transverse (x) direction. The self-modulation of the proton bunch is measured in imaging stations 1 and 2 and the optical and coherent transition radiation (OTR, CTR) diagnostics. The

central propagation axis by transverse electric fields that are present only when the proton bunch undergoes modulation in the plasma.

Electron bunches with a charge of 656 ± 14 pC (where the uncertainty is the r.m.s.) are produced and accelerated to 18.84 ± 0.05 MeV (where the uncertainty is the standard error of the mean) in a radio-frequency structure upstream of the vapour source³². These electrons are then transported along a beam line before being injected into the vapour source. Magnets along the beam line are used to control the injection angle and focal point of the electrons. For the results presented here, the electrons enter the plasma with a small vertical offset with respect to the proton bunch and a 200-ps delay with respect to the ionizing laser pulse (Fig. 1, bottom left). The beams cross approximately 2 m into the vapour source at a crossing angle of 1.2–2 mrad. Simulations show that electrons are captured in larger numbers and accelerated to higher energies when injected off-axis rather than collinearly with the proton bunch¹⁷. The normalized emittance of the witness electron beam at injection is approximately 11–14 mm mrad and its focal point is close to the entrance of the vapour source. The delay of 200 ps corresponds to approximately 25 proton microbunches resonantly driving the wakefield at $n_{pe} = 2 \times 10^{14}$ cm⁻³ and 50 microbunches at $n_{pe} = 7 \times 10^{14}$ cm⁻³.

A magnetic electron spectrometer (Fig. 1, right) enables measurement of the accelerated electron bunch³³. Two quadrupole magnets are located 4.48 m and 4.98 m downstream of the exit iris of the vapour source and focus the witness beam vertically and horizontally, respectively, to more easily identify a signal. These are followed by a 1-m-long C-shaped electromagnetic dipole with a maximum magnetic field of

approximately 1.4 T. A large triangular vacuum chamber sits in the cavity of the dipole. This chamber is designed to keep accelerated electron bunches under vacuum while the magnetic field of the dipole induces an energy-dependent horizontal deflection in the bunch. Electrons within a specific energy range then exit this vacuum chamber through a 2-mm-thick aluminium window and are incident on a 0.5-mm-thick gadolinium oxysulfide (Gd₂O₂S:Tb) scintillator screen (Fig. 1; blue, right) attached to the exterior surface of the vacuum chamber. The proton bunch is not greatly affected by the spectrometer magnets, owing to its high momentum, and continues to the beam dump. The scintillating screen is 997 mm wide and 62 mm high with semi-circular ends. Light emitted from the scintillator screen is transported over a distance of 17 m via three highly reflective optical-grade mirrors to an intensified charge-coupled device (CCD) camera fitted with a lens with a focal length of 400 mm. The camera and the final mirror of this optical line are housed in a dark room, which reduces ambient light incident on the camera to negligible values.

The energy of the accelerated electrons is inferred from their horizontal position in the plane of the scintillator. The relationship between this position and the energy of the electron is dependent on the strength of the dipole, which can be varied from approximately 0.1 T to 1.4 T. This position–energy relationship has been simulated using the Beam Delivery Simulation (BDSIM) code³⁴. The simulation tracks electrons of various energies through the spectrometer using measured and simulated magnetic-field maps for the spectrometer dipole, as well as the relevant distances between components. The accuracy of the magnetic-field maps, the precision of the distance measurements

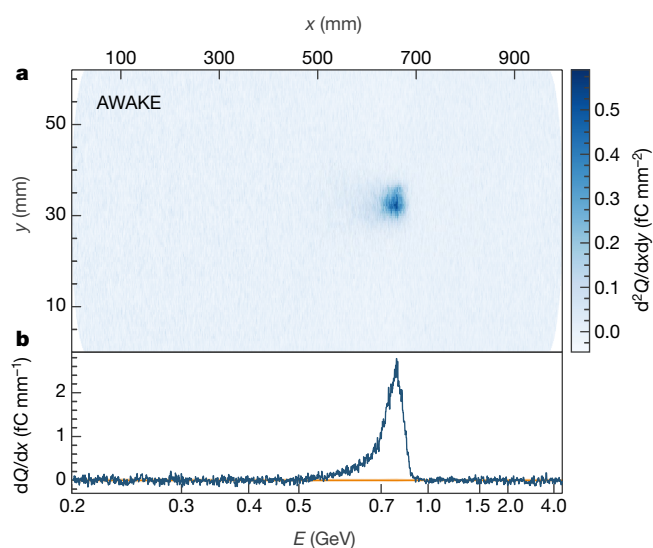


Fig. 2 | Signal of accelerated electrons. **a**, An image of the scintillator (with horizontal distance x and vertical distance y) with background subtraction and geometric corrections applied is shown, with an electron signal clearly visible. The intensity of the image is given in charge Q per unit area ($d^2Q/dxdy$), calculated using the central value from the calibration of the scintillator. **b**, A projection of the image in **a** is obtained by integrating vertically over the charge observed in the central region of the image. A 1σ uncertainty band from the background subtraction is shown in orange around zero. Both the image (**a**) and the projection (**b**) are binned in space, as shown on the top axis, but the central value from the position–energy conversion is indicated at various points on the bottom axis. The electron signal is clearly visible above the noise, with a peak intensity at an energy of $E \approx 800$ MeV.

and the 1.5-mm resolution of the optical system lead to an energy uncertainty of approximately 2%. The overall uncertainty, however, is dominated by the emittance of the accelerated electrons, and can be larger than 10%. The use of the focusing quadrupoles limits this uncertainty to approximately 5% for electrons near to the focused energy.

Owing to the difficulty of propagating an electron beam of well-known intensity to the spectrometer at AWAKE, the charge response of the scintillator is calculated using data acquired at CERN's Linear Electron Accelerator for Research (CLEAR) facility. This calibration is performed by placing the scintillator and vacuum window next to a beam charge monitor on the CLEAR beam line and measuring the scintillator signal. The response of the scintillator is found to depend linearly on charge over the range 1–50 pC. The response is also found to be independent of position and of energies in the range 100–180 MeV, to within the measurement uncertainty. This charge response is then recalculated for the optical system of the spectrometer at AWAKE by imaging a well-known light source at both locations. A response of $(6.9 \pm 2.1) \times 10^6$ CCD counts per incident picocoulomb of charge, given the acquisition settings used at AWAKE, is determined. The large 1σ uncertainty is due to different triggering conditions at CLEAR and AWAKE and systematic uncertainties in the calibration results.

Reliable acceleration of electrons relies on reproducible self-modulation of the proton beam. As well as the observation of the transverse expansion of the proton bunch, the optical and coherent transition radiation diagnostics showed clear microbunching of the beam. The proton microbunches were observed to be separated by the plasma wavelength (inferred from the measured rubidium vapour density, see Methods) for all parameter ranges investigated; they were also reproducible and stable in phase relative to the seeding. The detailed study of the self-modulation process will be the subject of separate AWAKE publications.

The data presented here were collected in May 2018. In Fig. 2a we show an image of the scintillator from an electron acceleration event at a plasma density of $1.8 \times 10^{14} \text{ cm}^{-3}$, with a measured density difference of $+5.3\% \pm 0.3\%$ over 10 m in the direction of propagation of the

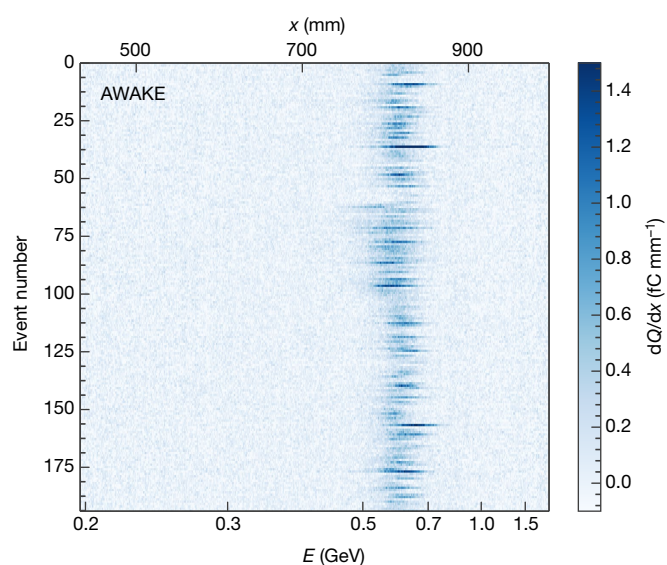


Fig. 3 | Background-subtracted projections of consecutive electron-injection events. Each projection (event) is a vertical integration over the central region of a background-subtracted spectrometer camera image. Brighter colours indicate regions of high charge density dQ/dx , corresponding to accelerated electrons. The quadrupoles of the spectrometer were varied to focus at energies of 460–620 MeV over the duration of the dataset. No other parameters were varied deliberately. The consistent peak around energy $E \approx 600$ MeV demonstrates the stability and reliability of the electron acceleration.

proton bunch. This image has been background-subtracted and corrected for vignetting and electron-angle effects (Methods). The quadrupoles of the spectrometer were focusing at an energy of approximately 700 MeV during this event, creating a substantial reduction in the vertical spread of the beam. In Fig. 2b we show a projection obtained by integrating over a central region of the scintillator. A 1σ uncertainty band, which comes from the background subtraction, is shown around zero. The peak in this figure has a high signal-to-noise ratio, which provides clear evidence of accelerated electrons. In both the image and the projection, the charge density is calculated using the central value of 6.9×10^6 CCD counts per picocoulomb. The asymmetric shape of the peak is due to the nonlinear position–energy relationship induced in the electron bunch by the magnetic field; when re-binned in energy, the signal peak is approximately Gaussian. Accounting for the systematic uncertainties described earlier, the observed peak has a mean of 800 ± 40 MeV, a FWHM of 137.3 ± 13.7 MeV and a total charge of 0.249 ± 0.074 pC. The amount of charge captured is expected to increase considerably¹⁷ as the emittance of the injected electron bunch is reduced and its geometric overlap with the wakefield is improved.

The stability and reliability of the electron acceleration is evidenced by Fig. 3, which shows projections from many consecutive electron-injection events. Each row in this plot is the background-subtracted projection from a single event, with the colour representing the signal intensity. The events correspond to a 2-h running period during which the quadrupoles were varied to focus over a range of approximately 460–620 MeV. Other parameters, such as the proton-bunch population, were not deliberately changed but vary naturally on a shot-to-shot basis. Despite the quadrupole scan and the natural fluctuations in the beam parameters, the plot still shows consistent and reproducible acceleration of electron bunches to approximately 600 MeV. The plasma density for these events is $1.8 \times 10^{14} \text{ cm}^{-3}$, with no density gradient. This lack of gradient is the cause of the difference in energy between the event in Fig. 2 and the events in Fig. 3.

The energy gain achievable by introducing a more optimal gradient is demonstrated in Fig. 4, which shows the peak energy achieved at different plasma densities with and without a gradient. The density gradients chosen are those that are observed to maximize the peak energy

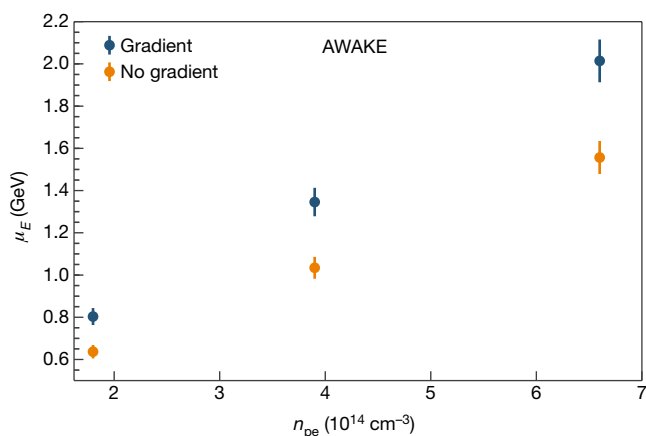


Fig. 4 | Measurement of the highest peak energies μ_E achieved at different plasma densities n_{pe} , with and without a gradient in the plasma density. The error bars arise from the position–energy conversion. The gradients chosen are those that were observed to maximize the energy gain. Acceleration to 2.0 ± 0.1 GeV is achieved with a plasma density of $6.6 \times 10^{14} \text{ cm}^{-3}$ with a density difference of $+2.2 \pm 0.1\%$ over 10 m.

for a given plasma density. At $1.8 \times 10^{14} \text{ cm}^{-3}$ the density difference was approximately $+5.3 \pm 0.3\%$ over 10 m, whereas at $3.9 \times 10^{14} \text{ cm}^{-3}$ and $6.6 \times 10^{14} \text{ cm}^{-3}$ it fell to $+2.5 \pm 0.3\%$ and $+2.2 \pm 0.1\%$, respectively. Given the precise control of the longitudinal plasma density, small density gradients can have a substantial effect on the acceleration because the electrons are injected tens of microbunches behind the ionizing laser pulse²⁶. The charge of the observed electron bunches decreases at higher plasma densities, owing in part to the smaller transverse size of the wakefield. In addition, the quadrupoles of the spectrometer have a maximum focusing energy of 1.3 GeV, which makes bunches accelerated to higher energies than this harder to detect above the background noise.

The energies shown in Fig. 4 are determined by binning the pixel data in energy and fitting a Gaussian over the electron signal region; the peak energy μ_E is the mean of this Gaussian. The observed energy spread of each bunch is determined by the width of this Gaussian and is approximately 10% of the peak energy. The peak energy increases with density, reaching 2.0 ± 0.1 GeV for $n_{pe} = 6.6 \times 10^{14} \text{ cm}^{-3}$ in the presence of a density gradient, at which point the charge capture is much lower. The energies of the accelerated electrons are within the range of values originally predicted by particle-in-cell and fluid code simulations of the AWAKE experiment^{17,18,26}. Future data-collection runs will address the effect of the electron-bunch delay, injection angle and other parameters on the accelerated energy and charge capture. These studies will help to determine what sets the limit on the energy gain.

In summary, we have demonstrated proton-driven plasma wakefield acceleration. The strong electric fields, generated by a series of proton microbunches, were sampled with a bunch of electrons. These electrons were accelerated up to 2 GeV in approximately 10 m of plasma and measured using a magnetic spectrometer. This technique has the potential to accelerate electrons to the teraelectronvolt scale in a single accelerating stage. Although still in the early stages of its programme, the AWAKE experiment is an important step towards realizing new high-energy particle physics experiments.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0485-4>.

Received: 22 June 2018; Accepted: 14 August 2018;

Published online 29 August 2018.

1. Tajima, T. & Dawson, J. M. Laser electron accelerator. *Phys. Rev. Lett.* **43**, 267–270 (1979).

2. Chen, P., Dawson, J. M., Huff, R. W. & Katsouleas, T. C. Acceleration of electrons by the interaction of a bunched electron beam with a plasma. *Phys. Rev. Lett.* **54**, 693–696 (1985); erratum **55**, 1537 (1985).
3. Joshi, C. The development of laser- and beam-driven plasma accelerators as an experimental field. *Phys. Plasmas* **14**, 055501 (2007).
4. Esarey, E., Schroeder, C. B. & Leemans, W. P. Physics of laser-driven plasma-based electron accelerators. *Rev. Mod. Phys.* **81**, 1229–1285 (2009).
5. Hogan, M. J. Electron and positron beam-driven plasma acceleration. *Rev. Accel. Sci. Technol.* **9**, 63–83 (2016).
6. Modena, A. et al. Electron acceleration from the breaking of relativistic plasma waves. *Nature* **377**, 606–608 (1995).
7. Mangles, S. P. D. et al. Monoenergetic beams of relativistic electrons from intense laser–plasma interactions. *Nature* **431**, 535–538 (2004).
8. Geddes, C. G. R. et al. High-quality electron beams from a laser wakefield accelerator using plasma-channel guiding. *Nature* **431**, 538–541 (2004).
9. Faure, J. et al. A laser–plasma accelerator producing monoenergetic electron beams. *Nature* **431**, 541–544 (2004).
10. Blumenfeld, I. et al. Energy doubling of 42 GeV electrons in a metre-scale plasma wakefield accelerator. *Nature* **445**, 741–744 (2007).
11. Litos, M. et al. High-efficiency acceleration of an electron beam in a plasma wakefield accelerator. *Nature* **515**, 92–95 (2014).
12. Schroeder, C. B., Esarey, E., Geddes, C. G. R., Benedetti, C. & Leemans, W. P. Physics considerations for laser–plasma linear colliders. *Phys. Rev. Spec. Top. Accel. Beams* **13**, 101301 (2010).
13. Caldwell, A., Lotov, K., Pukhov, A. & Simon, F. Proton-driven plasma-wakefield acceleration. *Nat. Phys.* **5**, 363–367 (2009).
14. Kumar, N., Pukhov, A. & Lotov, K. Self-modulation instability of a long proton bunch in plasmas. *Phys. Rev. Lett.* **104**, 255003 (2010).
15. Schroeder, C. B., Benedetti, C., Esarey, E., Grüner, F. J. & Leemans, W. P. Growth and phase velocity of self-modulated beam-driven plasma waves. *Phys. Rev. Lett.* **107**, 145002 (2011).
16. Pukhov, A. et al. Phase velocity and particle injection in a self-modulated proton-driven plasma wakefield accelerator. *Phys. Rev. Lett.* **107**, 145003 (2011).
17. Caldwell, A. et al. Path to AWAKE: evolution of the concept. *Nucl. Instrum. Methods A* **829**, 3–16 (2016).
18. Gschwendtner, E. et al. AWAKE, the advanced proton driven plasma wake field acceleration experiment at CERN. *Nucl. Instrum. Methods A* **829**, 76–82 (2016).
19. Muggli, P. et al. AWAKE readiness for the study of the seeded self-modulation of a 400 GeV proton bunch. *Plasma Phys. Control. Fusion* **60**, 014046 (2017).
20. Caldwell, A. & Lotov, K. V. Plasma wakefield acceleration with a modulated proton bunch. *Phys. Plasmas* **18**, 103101 (2011).
21. Caldwell, A. & Wing, M. VHEEP: a very high energy electron–proton collider. *Eur. Phys. J. C* **76**, 463–472 (2016).
22. Xia, G. et al. Collider design issues based on proton-driven plasma wakefield acceleration. *Nucl. Instrum. Methods A* **740**, 173–179 (2014).
23. Öz, E. & Muggli, P. A novel Rb vapor plasma source for plasma wakefield accelerators. *Nucl. Instrum. Methods A* **740**, 197–202 (2014).
24. Plyushchev, G., Kersevan, R., Petrenko, A. & Muggli, P. A rubidium vapor source for a plasma source for AWAKE. *J. Phys. D* **51**, 025203 (2018).
25. Fedosseev, V. N. et al. Integration of a terawatt laser at the CERN SPS beam for the AWAKE experiment on proton-driven plasma wake acceleration. In *Proc. 7th International Particle Accelerator Conference 2592–2595* (JACoW, 2016).
26. Petrenko, A., Lotov, K. & Sosedkin, A. Numerical studies of electron acceleration behind self-modulating proton beam in plasma with a density gradient. *Nucl. Instrum. Methods A* **829**, 63–66 (2016).
27. Sprangle, P. et al. Wakefield generation and GeV acceleration in tapered plasma channels. *Phys. Rev. E* **63**, 056405 (2001).
28. Lotov, K. V. Physics of beam self-modulation in plasma wakefield accelerators. *Phys. Plasmas* **22**, 103110 (2015).
29. Batsch, F. et al. Interferometer-based high-accuracy white light measurement of neutral rubidium density and gradient at AWAKE. *Nucl. Instrum. Methods A* <https://doi.org/10.1016/j.nima.2018.02.067> (2018).
30. Muggli, P. et al. Measuring the self-modulation instability of electron and positron bunches in plasmas. In *Proc. 6th International Particle Accelerator Conference 2506–2508* (JACoW, 2015).
31. Turner, M. et al. The two-screen measurement setup to indirectly measure proton beam self-modulation in AWAKE. *Nucl. Instrum. Methods A* **854**, 100–106 (2017).
32. Pepitone, K. et al. The electron accelerators for the AWAKE experiment at CERN—baseline and future developments. *Nucl. Instrum. Methods A* <https://doi.org/10.1016/j.nima.2018.02.044> (2018).
33. Deacon, L. et al. Development of a spectrometer for proton driven plasma wakefield accelerated electrons at AWAKE. In *Proc. 6th International Particle Accelerator Conference 2601–2604* (JACoW, 2015).
34. Nevay, L. et al. BDSIM: an accelerator tracking code with particle–matter interactions. Preprint at <https://arxiv.org/abs/1808.10745> (2018).

Acknowledgements All authors are members of the AWAKE Collaboration. This work was supported in part by: a Leverhulme Trust Research Project Grant RPG-2017-143 and by STFC (AWAKE-UK, Cockcroft Institute core and UCL consolidated grants), UK; the Russian Science Foundation (project number 14-50-00080) for simulations of oblique injection performed by Budker INP group; a Deutsche Forschungsgemeinschaft project grant PU 213-6/1 ‘Three-dimensional quasi-static simulations of beam self-modulation for plasma wakefield acceleration’; the National Research Foundation of Korea

(numbers NRF-2015R1D1A1A01061074 and NRF-2016R1A5A1013277); the Portuguese FCT—Foundation for Science and Technology, through grants CERN/FIS-TEC/0032/2017, PTDC-FIS-PLA-2940-2014, UID/FIS/50010/2013 and SFRH/IF/01635/2015; NSERC and CNRC for TRIUMF's contribution; and the Research Council of Norway. M. Wing acknowledges the support of the Alexander von Humboldt Stiftung and DESY, Hamburg. For their advice and contributions to the development of the magnetic spectrometer, we acknowledge B. Biskup, P. La Penna and M. Quattri. A. Petrenko acknowledges G. Demeter (Wigner Institute, Budapest) for calculating the rubidium ionization probability at AWAKE. F. Keeble acknowledges the operators of the CLEAR facility for their assistance during the calibration of the spectrometer. The AWAKE collaboration acknowledge the SPS team for proton delivery.

Reviewer information *Nature* thanks T. Tajima and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors contributed extensively to the work presented in this paper.

Competing interests The authors declare no competing interests.

Additional information

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.W.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

METHODS

Plasma generation. A CentAurus Ti:sapphire laser system is used to ionize the rubidium in the vapour source. The rubidium is confined by expansion chambers at the ends of the source with 10-mm-diameter irises through which rubidium flows constantly and condensates on the expansion walls. By the relation $\lambda_{pe} = 2\pi c[\epsilon_0 m_e / (n_{pe} e^2)]^{1/2}$, where c is the speed of light, ϵ_0 is the permittivity of free space, m_e is the electron mass and e is the electron charge, the available density range of $n_{pe} = 10^{14} - 10^{15} \text{ cm}^{-3}$ corresponds to a plasma wavelength of $\lambda_{pe} \approx 1.1 - 3.3 \text{ mm}$. The uniformity of the vapour density is ensured by flowing a heat-exchanging fluid around a concentric tube surrounding the source at a temperature stabilized to $\pm 0.05^\circ\text{C}$. Longitudinal density differences of between -10% and $+10\%$ over 10 m may be implemented, and controlled at the 1% level. The motion of the (heavy) rubidium ions can be neglected during the transit of the proton bunch because they are singly ionized³⁵.

Witness electron beam. Production of the witness electron beam is initiated by illuminating a Cs₂Te cathode by using a frequency-tripled laser pulse derived from the ionizing laser. Electron bunches with a charge of $656 \pm 14 \text{ pC}$ are produced and accelerated to an energy of 5.5 MeV in a 2.5 cell radio-frequency gun and are subsequently accelerated up to $18.84 \pm 0.05 \text{ MeV}$ using a 30 cell travelling wave structure. These electrons are then transported along an 18-m beam line before being injected into the vapour source. The focal point and crossing angle of the witness beam can be controlled via a combination of quadrupole and kicker magnets along this beam line.

Background subtraction. The large distance between the camera and the proton beam line means that background noise generated by radiation directly incident on the CCD is minimal. The scintillator of the spectrometer, however, is subject to considerable background radiation. The rise and decay of the scintillator signal occur on timescales longer than $1 \mu\text{s}$ and, as such, the scintillator photons captured by the camera are produced by an indivisible combination of background radiation and accelerated electrons. The majority of this background radiation is due to the passage of the proton bunch and comes from two main sources: a 0.2-mm-thick aluminium window located 43 m upstream of the spectrometer between AWAKE and the SPS transfer line, and a 0.6-mm-thick aluminium iris at the downstream end of the vapour source. The inner radius of this iris is 5 mm, leading to negligible interaction with the standard SPS proton bunch. However, protons that are defocused during self-modulation, such as those measured at the downstream imaging station, can interact with the iris, creating a substantial background. The strength of the transverse fields in the plasma and hence the number of protons that are defocused is strongly dependent on the plasma density. Consequently, the background generated by the defocused protons is more substantial at higher plasma densities, such as the AWAKE baseline density of $7 \times 10^{14} \text{ cm}^{-3}$. At this density, the radiative flux on the scintillator due to the iris is much higher than that from the thin window. Conversely, at a lower plasma density, such as $2 \times 10^{14} \text{ cm}^{-3}$, the radiation from the iris disappears completely and the remaining incident radiation is produced almost entirely by the interaction of the protons with the upstream window.

Owing to the variable nature of the radiation incident on the scintillator, background subtraction is a multistep process. A background data sample with the electron beam off at a plasma density of $1.8 \times 10^{14} \text{ cm}^{-3}$ is taken, such that the background has two key components: one due to the camera readout and ambient light in the experimental area, and another, N_p -dependent background caused by the proton bunch passing through the thin window. For each pixel imaging the scintillator, a linear function of N_p is defined by a χ^2 minimization fit to the background data sample, giving an N_p -dependent mean background image. For each signal event, a region of the scintillator is chosen where no accelerated electrons are expected, typically the lowest-energy part, and the background is rescaled by the ratio of the sums over this region in the signal event and the N_p -scaled background image. At higher plasma densities, a further step is needed to subtract the background from the iris. This background falls rapidly with increasing distance from the beam line and therefore depends on the horizontal position in the plane of the scintillator. A new region where the expected number of accelerated electrons is small is chosen, this time along the top and bottom edges of the scintillator. The mean of each column of pixels in this region is calculated and then subtracted from each pixel in the central region of that same column, leaving only the signal. The semi-circular ends of the scintillator reduce the effectiveness of this technique at the highest and lowest energies.

Signal extraction. To obtain an accurate estimate of the electron-bunch charge, the background-subtracted signal is corrected for two effects that vary across the horizontal plane of the scintillator. One effect comes from the variation in the horizontal angle of incidence of the electron on the scintillator. This angle is determined by the same tracking simulation used to define the position-energy relationship, and introduces a cosine correction to the signal owing to the variation in the path length of the electron through the scintillator. The second effect is vignetting, which occurs as result of the finite size of the optics of the spectrometer and the angular emission profile of the scintillator photons. A lamp that mimics this emission profile is scanned across the horizontal plane of the scintillator and the vignetting correction is determined by measuring its relative brightness. The increase in radiation accompanying the electron bunch, owing to its longer path length through the vacuum window at larger incident angles, is negligible and therefore does not require an additional correction factor.

Data reporting. No statistical methods were used to predetermine sample size.

Data availability

The datasets generated and analysed during this study are available from the corresponding author on reasonable request. The software code used in the analysis and to produce Figs. 2–4 is available from the corresponding author on reasonable request.

35. Vieira, J., Fonseca, R. A., Mori, W. B. & Silva, L. O. The ion motion in self-modulated plasma wakefield accelerators. *Phys. Rev. Lett.* **109**, 145005 (2012).

Deterministic teleportation of a quantum gate between two logical qubits

Kevin S. Chou^{1,2*}, Jacob Z. Blumoff^{1,2,3}, Christopher S. Wang^{1,2}, Philip C. Reinhold^{1,2}, Christopher J. Axline^{1,2}, Yvonne Y. Gao^{1,2}, L. Frunzio^{1,2}, M. H. Devoret^{1,2}, Liang Jiang^{1,2} & R. J. Schoelkopf^{1,2*}

A quantum computer has the potential to efficiently solve problems that are intractable for classical computers. However, constructing a large-scale quantum processor is challenging because of the errors and noise that are inherent in real-world quantum systems. One approach to addressing this challenge is to utilize modularity—a strategy used frequently in nature and engineering to build complex systems robustly. Such an approach manages complexity and uncertainty by assembling small, specialized components into a larger architecture. These considerations have motivated the development of a quantum modular architecture, in which separate quantum systems are connected into a quantum network via communication channels^{1,2}. In this architecture, an essential tool for universal quantum computation is the teleportation of an entangling quantum gate^{3–5}, but such teleportation has hitherto not been realized as a deterministic operation. Here we experimentally demonstrate the teleportation of a controlled-NOT (CNOT) gate, which we make deterministic by using real-time adaptive control. In addition, we take a crucial step towards implementing robust, error-correctable modules by enacting the gate between two logical qubits, encoding quantum information redundantly in the states of superconducting cavities⁶. By using such an error-correctable encoding, our teleported gate achieves a process fidelity of 79 per cent. Teleported gates have implications for fault-tolerant quantum computation³, and when realized within a network can have broad applications in quantum communication, metrology and simulations^{1,2,7}. Our results illustrate a compelling approach for implementing multi-qubit operations on logical qubits and, if integrated with quantum error-correction protocols, indicate a promising path towards fault-tolerant quantum computation using a modular architecture.

A quantum modular architecture is a distributed network (Fig. 1a) of modules that communicate with one another through quantum and classical channels. Each module is a small quantum processor that is composed of two separately optimized subsystems (Fig. 1b): data qubits that store and process quantum information and are realized as quantum memories; and communication qubits that mediate interactions between different modules. Each module operates individually as a highly functional node, capable of performing intra-module operations between the data and communication qubit subsystems. Inter-module operations between the data qubits are enabled by distributing entanglement between communication qubits. By adopting this modular approach, the data qubit subsystems are well-isolated from the other modules, which provides a systematic strategy for minimizing crosstalk and residual interactions across the entire network even when the system is scaled up. So far, elementary quantum networks have demonstrated the transmission of quantum information and the generation of entanglement between communication qubits^{8–11}. To extend the computational capabilities of these networks for universal quantum computation, it will be necessary to implement entangling operations on data qubits across different modules.

Owing to the inherent isolation between modules, multi-qubit operations in the modular architecture cannot be implemented using conventional approaches that rely on direct interactions, but instead utilize quantum teleportation^{3,12}. State teleportation was initially proposed¹² and later experimentally demonstrated^{13–19} as a technique for transferring an unknown quantum state between two quantum systems without transmitting the physical system that encodes the quantum state. This protocol relies on shared entanglement as a resource, along with local operations and classical communication between the two systems. Together these elements form the distinguishing characteristics of teleportation-based protocols in which information is transmitted through distinct quantum and classical channels. Expanding on this technique, the teleportation of a two-qubit quantum gate implements a unitary operation between two unknown states with a protocol that obviates the need for any direct interaction between the two data qubits (Fig. 1c)^{3–5,20}. Similar protocols have been demonstrated previously between two physical data qubits without real-time classical communication^{21–23}, with the desired operation extracted probabilistically through post-selection. However, to avoid excessive overhead and to make the modular approach scalable, it is crucial to perform these teleported gates deterministically.

In our work, we demonstrate a teleported CNOT gate that is deterministic and operates on logically encoded data qubits. A logical qubit is a two-dimensional subspace encoded within a higher-dimensional space designed with symmetry properties that allow for the detection and correction of certain errors. We implement two modules that each consist of a superconducting microwave cavity as the data qubit and a transmon as the communication qubit. Here, we generate entanglement between communication qubits via a local quantum bus that individually couples to each communication qubit. Our implementation can be adapted in the future to incorporate schemes for generating remote entanglement^{11,24}, which will be necessary for a scalable quantum modular architecture. We use a hardware-efficient approach^{6,25} to logically encode each data qubit within the states of a long-lived cavity mode. Importantly, despite the added complexity of our logical encoding, we implement high-fidelity control over both the data and communication qubit within each module. Using the teleported CNOT gate combined with real-time adaptive control, we generate a Bell state between two logical qubits and characterize the logical quantum process, thus validating our entangling operation on logical qubits.

Our physical implementation capitalizes on highly coherent and controllable elements from the three-dimensional circuit-quantum-electrodynamics platform. Each module (Fig. 1d, Methods) consists of a high-quality-factor (high- Q) three-dimensional electromagnetic cavity²⁶ as the data qubit, a transmon qubit as the communication qubit and a Purcell-filtered, low- Q stripline resonator²⁷ for readout of the transmon qubit. The transmon qubit is capacitively coupled to both the data qubit and the readout resonator. We achieve data qubit lifetimes (of about 1 ms) that are around three orders of magnitude greater than the measurement time (less than about 1 μ s), enabling

¹Department of Applied Physics and Physics, Yale University, New Haven, CT, USA. ²Yale Quantum Institute, Yale University, New Haven, CT, USA. ³Present address: HRL Laboratories, Malibu, CA, USA. *e-mail: kevin.chou@yale.edu; robert.schoelkopf@yale.edu

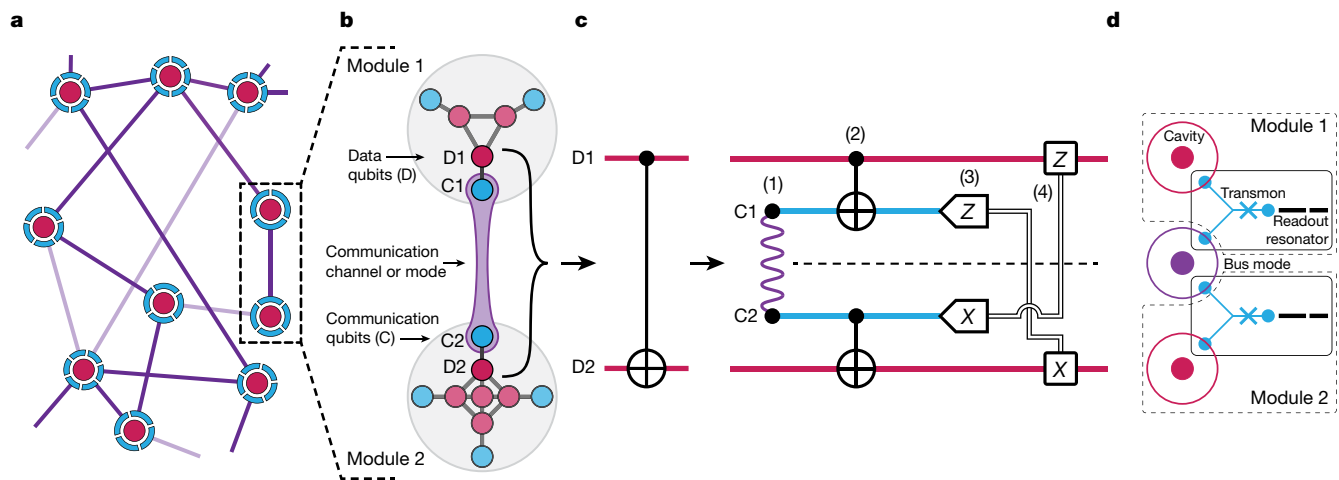


Fig. 1 | Construction of a modular architecture and teleported CNOT gate. **a**, Network overview of the modular quantum architecture. Modules are represented as nodes of a quantum network and are composed of data qubit(s) (magenta) and communication qubit(s) (cyan). Coupling between modules is generated through reconfigurable communication channels that may be enabled (dark purple lines) or disabled (light purple lines). **b**, Quantum modules. Each module houses a small quantum processor that is capable of high-fidelity operations among data qubits and communication qubits. In our experiment, we create two modules, each consisting of one data qubit (D1 and D2) and one communication qubit (C1 and C2). **c**, Teleported CNOT circuit between D1 and D2. The teleported CNOT circuit requires: (1) entanglement between C1 and C2

(purple meander), (2) local operations, (3) measurement of C1 in the \hat{Z} basis and C2 in the \hat{X} basis, where \hat{X} and \hat{Z} are Pauli operators, and (4) classical communication (double lines) and feedforward operations. **d**, Experimental realization (schematic top view) in a three-dimensional circuit-quantum-electrodynamics implementation. Each module consists of a data qubit defined as a coaxial quarter-wavelength ($\lambda/4$) three-dimensional cavity (magenta), a communication qubit defined as a Y-shaped transmon qubit (cyan) and a Purcell-filtered, quasi-planar, $\lambda/2$ stripline readout resonator (black). In this experiment, the two modules are linked by an additional mode realized as a coaxial $\lambda/4$ three-dimensional cavity (purple) that serves as a bus mode (Extended Data Fig. 1, Supplementary Information).

both quantum information storage and fast measurement within a single package (Supplementary Information). In this experiment, the communication channel is implemented as an additional cavity mode and functions as a quantum bus (hereafter, ‘bus’), coupling individually to both communication qubits. Although we utilize this local mode to link the two modules, the two data qubits have an immeasurably small direct coupling, which is bounded to be at least an order of magnitude smaller than the smallest decay rate in our system (Supplementary Information). Therefore, despite the physical proximity (around 2 cm) between the two modules, our two data qubits are effectively non-interacting, demonstrating the same isolation distinctive of remote modular architectures. In addition, our entire device exhibits low readout crosstalk, which is critical for the teleported gate and is another characteristic property of independent modules (Methods).

The high-dimensional cavity modes that define our data qubits allow for a wide range of encodings, including those that can address the dominant errors that face cavity memories^{6,25}. For our data qubits to fulfil the role of a quantum memory, we chose to encode each data qubit using the first-level bosonic binomial quantum code²⁵, which has logical basis states (specified in the photon-number basis of the cavity)

$$|0_L\rangle = |2\rangle, |1_L\rangle = \frac{|0\rangle + |4\rangle}{\sqrt{2}}$$

Our protocol is flexible to different data-qubit encodings, and we also present results using the $|0\rangle$ and $|1\rangle$ Fock basis, which requires only an in situ software modification to the control pulses. The first-level binomial encoding specifies a logical qubit that can be used to detect and correct for single-photon loss, the main error mechanism for cavity memories (Methods)⁶. Demonstration of the teleported gate using such a logical encoding offers a practical route for incorporating quantum error-correction protocols in the future. As an illustration of our control of the logical qubit, we prepared six cardinal states of the logical Bloch sphere and characterized each state by measuring the Wigner function of the data qubit (Fig. 2a). The Wigner function not only provides a

strikingly visual representation of the logical qubit state, but also completely specifies the underlying cavity state, a capability analogous to full state tomography of the constituent physical qubits that compose a logical qubit.

The teleported CNOT gate starts with the generation of entanglement in the communication qubits to create a communication channel between the two modules (step 1 in Fig. 1c). Any maximally entangled state is acceptable, and the specific choice requires only small modifications to later steps of the teleported-gate protocol. In our implementation, we use the Bell state $|\Psi^+\rangle = (|ge\rangle + |eg\rangle)/\sqrt{2}$, where $|g\rangle$ and $|e\rangle$ are the ground and first excited states, respectively, of the transmon qubits and specify the basis states of the communication qubits. The state is generated by performing a resonator-induced phase (RIP) gate²⁸ on the bus and single-qubit rotations on the communication qubits (Methods). Using this gate, we generated a Bell pair between the communication qubits in approximately 680 ns with a state fidelity of $97\% \pm 1\%$ as determined from quantum state tomography (Methods; the error quoted here and elsewhere is defined in Methods and Supplementary Information).

Next, local operations performed within each module entangle the data and communication qubits (step 2 in Fig. 1c). Our local operations are implemented using optimal-control techniques, which enable universal quantum control between the data and communication qubits²⁹. We generate all of our local operations with pulse lengths between 1 μ s and 2 μ s. Characterization of these logical operations yields single-data-qubit and two-qubit (between the data and communication qubits) gate fidelities of around 97% and 94%, respectively (Supplementary Information).

After the entangling local operations, we perform measurements on the communication qubits (step 3 in Fig. 1c), thereby effecting a unitary operation between only the two data qubits. It is essential that the measurements do not reveal information about the state of the data qubits. In the teleported-gate protocol, this is accomplished by individual measurements of the communication qubits in the \hat{Z} and \hat{X} bases (where \hat{Z} and \hat{X} are Pauli operators), which lead to four uniformly

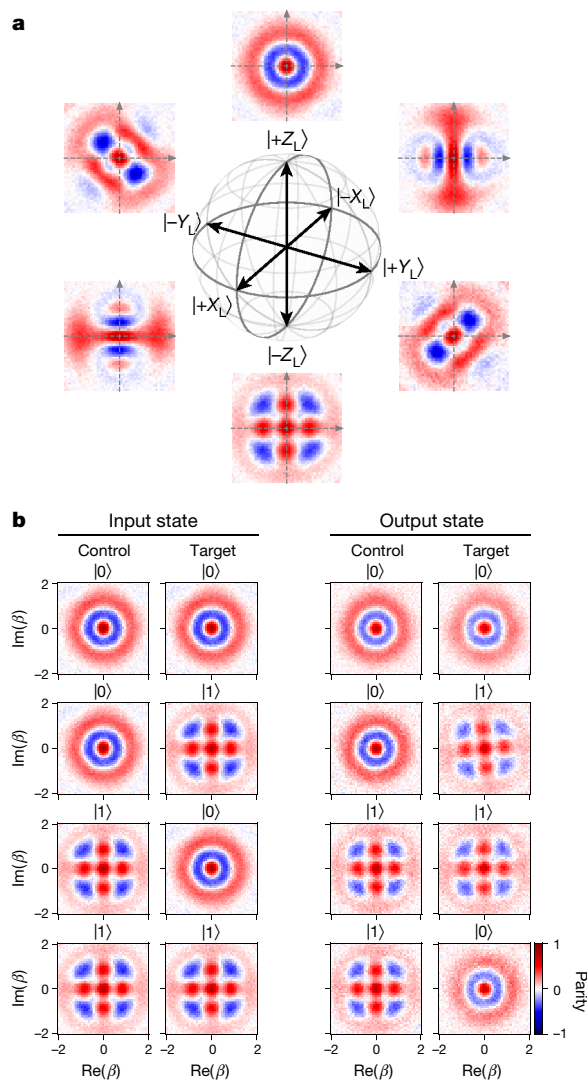


Fig. 2 | Logical data-qubit encoding and CNOT truth table. **a**, Logical Bloch sphere for the binomial code encoding. The data qubit is logically encoded in the binomial code basis and the Wigner function for each of the six cardinal states $\{\pm\hat{Z}_L, \pm\hat{X}_L, \pm\hat{Y}_L\}$ is shown (colour scale as in **b**). **b**, Teleported CNOT truth table. The left two columns show experimental Wigner functions for all four logical computational states as input states, and the right two columns show the extracted Wigner functions after performing the teleported CNOT operation, illustrating the correct classical behaviour of the gate. We determine the scaled Wigner function by directly measuring the displaced joint parity of the cavity, which is parameterized by β , a complex variable of the cavity state.

distributed outcomes. Each outcome heralds a unitary operation between the two data qubits that is a CNOT gate up to single-qubit operations. As a result, high-fidelity measurements are necessary to correctly determine the particular operation enacted on the data qubits. In our system, we achieve single-shot state-assignment fidelities of the communication qubits of around 99% (Methods).

Finally, ensuring that the protocol implements the desired CNOT operation independently of the measurement outcome requires classical communication and feedforward operations (step 4 in Fig. 1c). Two classical bits of information are needed to communicate measurement results between modules. This information is used to apply feedforward operations, transforming the protocol into a deterministic operation and thus completing the teleportation. In our experiment, the measurements must be non-destructive to the communication qubits because these qubits are used for subsequent steps of our protocol. To enable the measurements and the feedforward operations, we use a real-time controller⁶ to orchestrate quantum programs for our

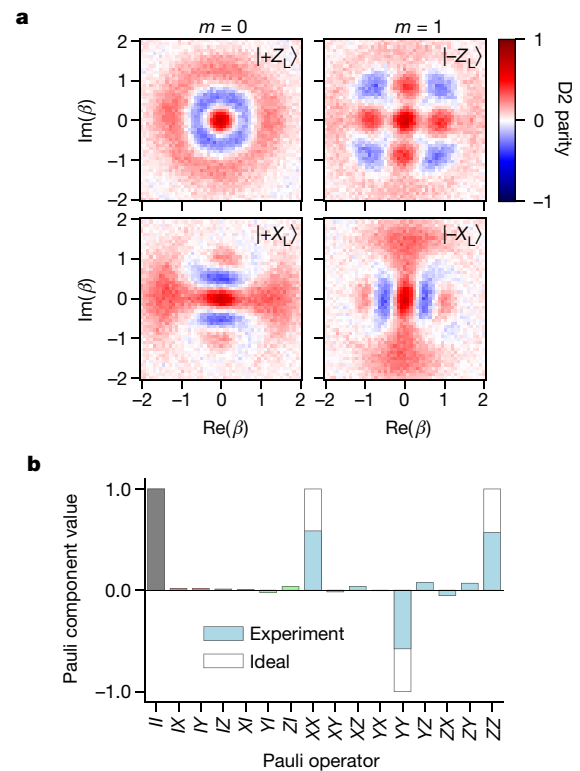


Fig. 3 | Generation of a logical Bell state. **a**, Quantum correlations of a logical Bell state $|\Phi_L^+\rangle = (|0_L 0_L\rangle + |1_L 1_L\rangle)/\sqrt{2}$ is first created using the teleported CNOT gate. The control qubit (D1) is measured in either the logical \hat{Z}_L basis (top) or \hat{X}_L basis (bottom), and Wigner tomography is performed on the target qubit (D2), conditioned on the measurement result m of the control qubit, $m=0$ (left) or $m=1$ (right). When measuring in the \hat{Z}_L basis (top), a measurement result of $m=0$ ($m=1$) indicates that the control qubit is found to be in $|+Z_L\rangle$ ($|-Z_L\rangle$) and the target qubit in $|+Z_L\rangle$ ($|-Z_L\rangle$). When measuring in the \hat{X}_L basis (bottom), a measurement result of $m=0$ ($m=1$) indicates that the control qubit is found to be in $|+X_L\rangle$ ($|-X_L\rangle$) and target qubit in $|+X_L\rangle$ ($|-X_L\rangle$). Correlations between the measurement result and the measured state indicate the generation of an entangled state between D1 and D2. **b**, Logical state tomography. After generating $|\Phi_L^+\rangle$, logical qubit tomography is performed on both the control and the target qubit. The reconstructed state, represented in the Pauli basis $\{I, X, Y, Z\}$, confirms that the teleported CNOT gate has generated the target Bell state, except for reduced contrast from the ideal value. The reconstructed states shown in blue correspond to two-qubit correlations, those in red and green to single-qubit correlations and that in grey to the identity.

experiment, combining control, measurement, state estimation and feedforward in a single integrated system. For every experimental run, this controller handles the distribution of classical information between the two modules and the application of the feedforward operations, all within a fraction of the lifetime (about 1%) of the communication qubits. We independently analysed the measurement and feedforward processes to have a combined fidelity of approximately 97%, excluding the conditional data-qubit operations (Methods).

Therefore, by consuming a shared entangled pair and communicating two classical bits of information, this procedure effects a CNOT operation between the data qubits without requiring a unitary operation between the two modules after the generation of the shared entangled pair. Having demonstrated all of the elements necessary for realizing the teleported CNOT gate, we characterized the full two-qubit gate through a series of four separate analyses.

In the first analysis, we verified the classical behaviour of the gate by generating a truth table for the set of computational states. We prepared the data qubits each of the four states $\{|0_L 0_L\rangle, |0_L 1_L\rangle, |1_L 0_L\rangle, |1_L 1_L\rangle\}$ and enacted the teleported CNOT on each, ideally leading to the output

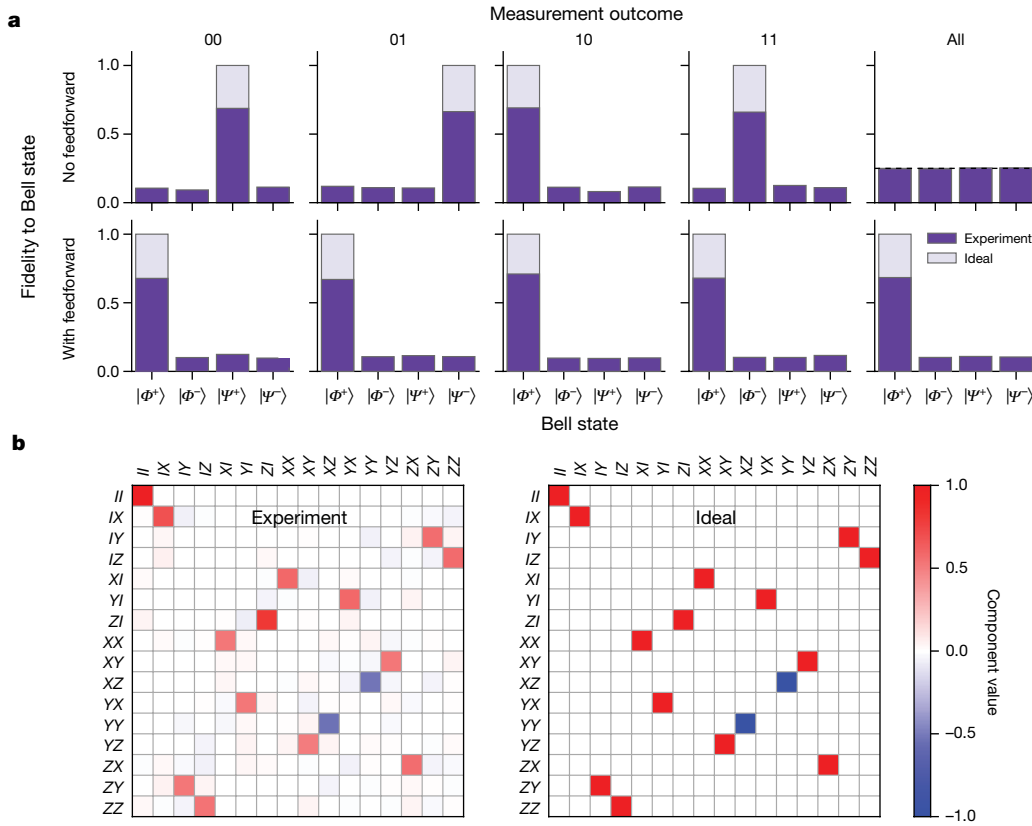


Fig. 4 | Demonstration of a deterministic teleported CNOT gate.

a, Effect of feedforward operations. The teleported CNOT gate is applied to the initial state $|\psi_{\text{init}}\rangle = (|0_L\rangle + |1_L\rangle)|0_L\rangle/\sqrt{2}$ and the fidelity of the resulting state to each of the four Bell states is extracted. When feedforward operations are not applied (top), each measurement outcome $\{00, 01, 10, 11\}$ results in a different Bell state. If all measurement results are compiled together, the resulting state is completely mixed ('All'). On the other hand, if the feedforward operations are applied (bottom), then

the correct state $|\Phi^+\rangle$ is found for every measurement outcome.

b, Quantum process tomography of the teleported CNOT gate. We represent the quantum process $\mathcal{R}_{\text{CNOT}}$ in the Pauli transfer representation, in which the process map is expressed in the Pauli basis: $\mathbf{P}_{\text{out}} = \mathcal{R}_{\text{CNOT}} \mathbf{P}_{\text{in}}$, given input- and output-state Pauli vectors $\mathbf{P}_{\text{in,out}}$ (Methods). Agreement between the experimentally reconstructed (left) and ideal (right) processes indicates the successful implementation of a deterministic teleported CNOT gate.

states $\{|0_L 0_L\rangle, |0_L 1_L\rangle, |1_L 1_L\rangle, |1_L 0_L\rangle\}$. We extracted the input and output states by measuring Wigner functions for each data qubit. Our results (Fig. 2b) provide qualitative validation of the teleported CNOT gate on the computational basis states.

In the second analysis, we demonstrated that our gate is a distinctly quantum operation by using the teleported CNOT gate to generate entanglement between two logical qubits. We prepared the data qubits in the separable initial state $|\psi_{\text{init}}\rangle = (|0_L\rangle + |1_L\rangle)|0_L\rangle/\sqrt{2}$ and performed the gate. The ideal output state is the Bell state $|\Phi_L^+\rangle = (|0_L 0_L\rangle + |1_L 1_L\rangle)/\sqrt{2}$. We verified that our teleported CNOT gate generates this logical-qubit Bell pair using two separate methods, which together highlight our ability to characterize the data qubits on a logical level (the encoded two-dimensional subspace) and on a physical level (the multi-dimensional cavity state).

In the first method, we performed a pair of experiments to show that the state exhibits quantum correlations. Given the target state $|\Phi_L^+\rangle$, when we measure the control qubit in the logical \hat{Z}_L basis and find it in $|0_L\rangle$ ($|1_L\rangle$), we expect the target qubit to be $|0_L\rangle$ ($|1_L\rangle$). We enacted the logical \hat{Z}_L measurement and, conditioned on the result, performed physical-qubit tomography on the target data qubit by measuring its Wigner function (Methods). As expected, we observed strong \hat{Z} correlations between the control and target data qubits (Fig. 3a, top). Next, we rotated the measurement basis and performed \hat{X}_L measurements of the control data qubit. Conditioned on the control data qubit in the state $|\pm X_L\rangle = (|0_L\rangle \pm |1_L\rangle)/\sqrt{2}$, we experimentally found the target data qubit to be in the expected state $|\pm X_L\rangle$ (Fig. 3a, bottom), thus establishing \hat{X} correlations between the two data qubits. These two complementary experiments confirm the non-classical nature of the

experimental logical Bell state and indicate that our gate produced a non-separable two-qubit state.

In the second method, we analysed the joint state within the logical subspace of the two data qubits by performing quantum state tomography (Methods). We reconstructed the two-qubit state in the Pauli basis (Fig. 3b), extracting a state fidelity of $\mathcal{F}_{\text{Bell}} = 68\% \pm 1\%$ and concurrence of $\mathcal{C} = 0.37 \pm 0.01$, which exceeds the threshold for a classically correlated state. These quantities include imperfections associated with logical-state preparation and decoding operations, which together contribute about 6% infidelity for each data qubit. Using the teleported CNOT gate, we have thus generated a Bell state between logical qubits encoded as multi-photon states that, from inspection of the reconstructed density operator, has dominant two-qubit correlations (for example, two-qubit parity $\langle ZZ \rangle = 0.57$) and near-zero single-qubit correlations (for example, the single-qubit Z correlations for D1 and D2 are $\langle ZI \rangle = 0.04$ and $\langle IZ \rangle = 0.01$, respectively).

Our implementation of the teleported gate as a deterministic operation requires reliable classical communication and feedforward operations. In the third analysis, we investigated the importance of classical communication by performing the previously described entanglement sequence, recording the measurement outcomes and extracting four conditioned output states. We performed this sequence with and without applying the feedforward operations (step 4). Each measurement outcome $\{00, 01, 10, 11\}$ ideally occurs with probability 1/4 and, without the feedforward operations, heralds one of four Bell states $\{|\Psi_L^+\rangle, |\Psi_L^-\rangle, |\Phi_L^+\rangle, |\Phi_L^-\rangle\}$, where $|\Psi_L^\pm\rangle = (|0_L 1_L\rangle \pm |1_L 0_L\rangle)/\sqrt{2}$ and $|\Phi_L^\pm\rangle = (|0_L 0_L\rangle \pm |1_L 1_L\rangle)/\sqrt{2}$. Our results (Fig. 4a; top, first four panels) are consistent with the ideal outcome, save for reduced contrast, and

we extracted conditioned fidelities of {69%, 66%, 69%, 66%} and outcome frequencies of {0.25, 0.26, 0.24, 0.25}. The fact that we generated different Bell pairs indicates that each conditional operation is a CNOT gate up to single-qubit operations. Without real-time knowledge of these measurement outcomes, these states will all add incoherently, resulting in a completely mixed state in which all information has been lost (Fig. 4a; top, 'All'). If we instead post-selected on the measurement outcomes, the operation would be left as a probabilistic two-qubit gate, achieving the target operation only 1/4 of the time (Fig. 4a; top, measurement outcome 10). Therefore, it is only when we combine real-time classical communication and feedforward that we can implement a deterministic teleported operation that performs the correct process for all measurement outcomes (Fig. 4a, bottom).

Finally, in the fourth analysis, we fully characterized the logical process for the teleported CNOT gate. We performed quantum process tomography on the two logical qubits and our reconstructed process agrees qualitatively with the expected process (Fig. 4b). From the experimental reconstruction, we calculate a process fidelity of $\mathcal{F}_{\text{pro}} = 68\% \pm 2\%$ without accounting for logical encoding or decoding steps that subtract from the extracted gate fidelity. With these corrections included (Supplementary Information), we infer a process fidelity of $\mathcal{F}_{\text{gate}} = 79\% \pm 2\%$ for our teleported CNOT gate. To evaluate the experimental performance of the teleported gate, we assembled an error budget that combines the infidelity of each element of the gate, accounting for the known imperfections of our system. From this analysis (Supplementary Information), we expect a gate fidelity of $\mathcal{F}_{\text{thy}} \approx 84\% \pm 3\%$, which is consistent with experimental results. This indicates that other non-idealities, such as residual interactions or imperfect system characterization, are smaller effects in our system. We also performed the teleported operation using the Fock $|0\rangle$ and $|1\rangle$ states, achieving a process fidelity of $\mathcal{F}_{\text{gate}} = 86\% \pm 2\%$ when accounting for the encoding and decoding steps. This difference in process fidelity is well understood: the binomial encoding has a higher average photon number ($\bar{n} = 2$) compared to the Fock encoding ($\bar{n} = 0.5$), which results in an additional overhead in the photon loss rate and increased complexity of the local operations. Although the fidelity of the binomial encoding is lower than that of the Fock encoding, the advantage of the binomial encoding is that a single-photon loss event can, in principle, be detected via parity measurements and is therefore a correctable error.

The performance of our teleported gate, although an encouraging step towards operations between logical qubits, requires improvement to be useful for the modular quantum architecture. An advantage of our work is that the teleported gate is modular and uses relatively modest elements, all of which are part of the standard toolbox for quantum computation in general. Therefore, on-going progress to improve any of the elements will directly increase gate performance. There already exist well-defined prescriptions to improve each element of the teleported gate. For example, a communication qubit implemented using a high-Q cavity instead of a transmon qubit would directly address the dominant source of infidelity in our implementation—the communication-qubit coherence time of $T_2 \approx 15 \mu\text{s}$. Such a modification would not only enable the development of a module containing error-correctable data and communication qubits, but also introduce opportunities to improve intra-module operations via multi-cavity gates³⁰ and communication-qubit measurements via robust, repeated readout strategies of a bosonic mode^{5,31}.

The protocol for the teleported CNOT gate used in this work is one example of an extensive family of two-qubit operations that may be implemented using the same resources^{3–5}. Such teleportation-based gates are important primitives for the implementation of a modular architecture and may be part of a broader approach to fault-tolerant quantum computation^{2,3,32}. One of the next steps will be to demonstrate non-local teleported gates using spatially separate modules, which will require remote entanglement. Because this entanglement can be prepared before the teleported operation, the gate is agnostic to how the entanglement is generated. Therefore, the protocol can take advantage of various approaches, including deterministic³³ and probabilistic³⁴ schemes, and should benefit from entanglement-purification

protocols^{5,35}. Building on our results and recent demonstrations of remote entanglement in circuit-quantum-electrodynamics systems^{11,24}, it should be possible to integrate these technologies, enabling the development of modular quantum computing using superconducting qubits.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0470-y>

Received: 5 January; Accepted: 27 June 2018;

Published online: 05 September 2018

- Kimble, H. J. The quantum internet. *Nature* **453**, 1023–1030 (2008).
- Monroe, C. et al. Large-scale modular quantum-computer architecture with atomic memory and photonic interconnects. *Phys. Rev. A* **89**, 022317 (2014).
- Gottesman, D. & Chuang, I. L. Demonstrating the viability of universal quantum computation using teleportation and single-qubit operations. *Nature* **402**, 390–393 (1999).
- Eisert, J., Jacobs, K., Papadopoulos, P. & Plenio, M. B. Optimal local implementation of nonlocal quantum gates. *Phys. Rev. A* **62**, 052317 (2000).
- Jiang, L., Taylor, J. M., Sørensen, A. S. & Lukin, M. D. Distributed quantum computation based on small quantum registers. *Phys. Rev. A* **76**, 062323 (2007).
- Ofek, N. et al. Extending the lifetime of a quantum bit with error correction in superconducting circuits. *Nature* **536**, 441–445 (2016).
- Duan, L.-M., Lukin, M. D., Cirac, J. I. & Zoller, P. Long-distance quantum communication with atomic ensembles and linear optics. *Nature* **414**, 413–418 (2001).
- Ritter, S. et al. An elementary quantum network of single atoms in optical cavities. *Nature* **484**, 195–200 (2012).
- Bernien, H. et al. Heralded entanglement between solid-state qubits separated by three metres. *Nature* **497**, 86–90 (2013).
- Hucul, D. et al. Modular entanglement of atomic qubits using photons and phonons. *Nat. Phys.* **11**, 37–42 (2015).
- Narla, A. et al. Robust concurrent remote entanglement between two superconducting qubits. *Phys. Rev. X* **6**, 031036 (2016).
- Bennett, C. H. et al. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.* **70**, 1895–1899 (1993).
- Bouwmeester, D. et al. Experimental quantum teleportation. *Nature* **390**, 575–579 (1997).
- Furusawa, A. et al. Unconditional quantum teleportation. *Science* **282**, 706–709 (1998).
- Riebe, M. et al. Deterministic quantum teleportation with atoms. *Nature* **429**, 734–737 (2004).
- Barrett, M. D. et al. Deterministic quantum teleportation of atomic qubits. *Nature* **429**, 737–739 (2004).
- Sherson, J. F. et al. Quantum teleportation between light and matter. *Nature* **443**, 557–560 (2006).
- Olmschenk, S. et al. Quantum teleportation between distant matter qubits. *Science* **323**, 486–489 (2009).
- Steffen, L. et al. Deterministic quantum teleportation with feed-forward in a solid state system. *Nature* **500**, 319–322 (2013).
- Gottesman, D. The Heisenberg representation of quantum computers. Preprint at <https://arxiv.org/abs/quant-ph/9807006> (1998).
- Huang, Y.-F., Ren, X.-F., Zhang, Y.-S., Duan, L.-M. & Guo, G.-C. Experimental teleportation of a quantum controlled-NOT gate. *Phys. Rev. Lett.* **93**, 240501 (2004).
- Gao, W.-B. et al. Teleportation-based realization of an optical quantum two-qubit entangling gate. *Proc. Natl Acad. Sci. USA* **107**, 20869–20874 (2010).
- K., V. P., Joy, D., Behera, B. K. & Panigrahi, P. K. Experimental demonstration of non-local controlled-unitary quantum gates using a five-qubit quantum computer. Preprint at <https://arxiv.org/abs/1709.05697> (2017).
- Roch, N. et al. Observation of measurement-induced entanglement and quantum trajectories of remote superconducting qubits. *Phys. Rev. Lett.* **112**, 170501 (2014).
- Michael, M. H. et al. New class of quantum error-correcting codes for a bosonic mode. *Phys. Rev. X* **6**, 031006 (2016).
- Reagor, M. et al. Quantum memory with millisecond coherence in circuit QED. *Phys. Rev. B* **94**, 014506 (2016).
- Axline, C. et al. An architecture for integrating planar and 3D cQED devices. *Appl. Phys. Lett.* **109**, 042601 (2016).
- Paik, H. et al. Experimental Demonstration of a Resonator-Induced Phase Gate in a Multiqubit Circuit-QED System. *Phys. Rev. Lett.* **117**, 250502 (2016).
- Heeres, R. W. et al. Implementing a universal gate set on a logical qubit encoded in an oscillator. *Nat. Commun.* **8**, 94 (2017).
- Rosenblum, S. et al. A CNOT gate between multiphoton qubits encoded in two cavities. *Nat. Commun.* **9**, 652 (2018).
- Hann, C. T. et al. Robust readout of bosonic qubits in the dispersive coupling regime. *Phys. Rev. A* **98**, 022305 (2018).
- Nickerson, N. H., Li, Y. & Benjamin, S. C. Topological quantum computing with a very noisy network and local error rates approaching one percent. *Nat. Commun.* **4**, 1756 (2013).
- Cirac, J. I., Zoller, P., Kimble, H. J. & Mabuchi, H. Quantum state transfer and entanglement distribution among distant nodes in a quantum network. *Phys. Rev. Lett.* **78**, 3221–3224 (1997).

34. Barrett, S. D. & Kok, P. Efficient high-fidelity quantum computation using matter qubits and linear optics. *Phys. Rev. A* **71**, 060310 (2005).
35. Bennett, C. H. et al. Purification of noisy entanglement and faithful teleportation via noisy channels. *Phys. Rev. Lett.* **76**, 722–725 (1996).

Acknowledgements We thank B. J. Lester, Z. K. Mineev, A. Narla, U. Vool and I. L. Chuang for discussions on the manuscript, and A. Narla, K. Sliwa and N. Frattini for assistance on the parametric amplifier. Facilities use was supported by the Yale SEAS cleanroom, YINQE and NSF MRSEC DMR-1119826. This research was supported by the Army Research Office under grant numbers W911NF-14-1-0011 and W911NF-16-10349 and by the Air Force Office of Scientific Research under grant numbers FA9550-14-1-0052 and FA9550-15-1-0015. C.J.A. acknowledges support from a NSF Graduate Research Fellowship under grant number DGE-1122492. Y.Y.G. was supported by an A*STAR NSS Fellowship. L.J. acknowledges additional support from the Alfred P. Sloan Foundation under grant number BR2013-049 and from the Packard Foundation under grant number 2013-39273.

Author contributions K.S.C., J.Z.B. and C.S.W. performed the experiment and analysed the data under the supervision of R.J.S. P.C.R. developed the

feedforward control software and implemented the software used to generate optimal-control pulses. C.J.A., Y.Y.G. and L.F. fabricated the transmon qubits. K.S.C., J.Z.B. and R.J.S. designed the experiment. L.J., M.H.D. and L.F. provided theoretical support. K.S.C. and R.J.S. wrote the manuscript with contributions from all authors.

Competing interests R.J.S., M.H.D. and L.F. are founders, and R.J.S. and L.F. are equity shareholders, of Quantum Circuits, Inc.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0470-y>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0470-y>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to K.S.C. and R.J.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Experimental device and setup. Our experiment uses three high- Q $\lambda/4$ coaxial three-dimensional cavities machined out of a single block of aluminium (99.99% purity) and two sapphire chips on which a Y-shaped transmon qubit and a quasi-planar $\lambda/2$ readout resonator are lithographically defined (Extended Data Fig. 1). The device is cooled to $T \approx 10$ mK in a dilution refrigerator. Our control hardware includes an FPGA-based controller that functions both as a real-time arbitrary-waveform generator and a digitizer for performing measurements of the system. These measurements are enabled through the use of two independent readout chains that each include a Josephson parametric converter (JPC) for nearly quantum limited amplification of output signals. More detailed information regarding device construction and cryogenic wiring is provided in Supplementary Information.

Transmon measurement. In this experiment, each module is connected to a separate JPC for fast, high-fidelity measurement of the transmon qubit. We achieve single-shot assignment fidelities of around 99.4%, largely limited by transmon decay during the measurement pulse of 600 ns. We define the assignment fidelity as the average probabilities of correctly assigning the state when we prepare the transmon in $|g\rangle$ and $|e\rangle$: $\mathcal{F}_{\text{assign}} = [\text{Pr}(g' | |g\rangle) + \text{Pr}(e' | |e\rangle)]/2$. This high-quality measurement, coupled with the real-time capabilities of our quantum controller, enables conditional operations based on an extracted measurement result. The length of time from the start of a measurement pulse to the application of a conditioned operation is around 1,000 ns, which includes the length of the measurement pulse (600 ns), cable delays (200 ns), and integration and state-estimation latencies (200 ns).

It is critical that the two communication qubit measurements be independent for the demonstration of the teleported gate. To assess the measurement crosstalk, we perform a Rabi experiment and simultaneous measurements on both communication qubits (Extended Data Fig. 2a). Our results (Extended Data Fig. 2b, c) indicate that the measurements are highly selective to the qubit addressed. From our data, we estimate the measurement crosstalk—defined to be the ratio of the measurement contrast of measuring the directly coupled qubit to that of measuring the isolated qubit—to be less than 10^{-4} . In future implementations of this experiment in which the two modules are physically separate, the measurement crosstalk will be completely negligible.

Data-qubit encodings. Binomial encoding. As discussed in the main text, we demonstrate the teleported gate using one of the binomial quantum codes²⁵, with basis states $|0_L\rangle = |2\rangle$ and $|1_L\rangle = (|0\rangle + |4\rangle)/\sqrt{2}$. This logical encoding provides the ability to perform quantum error correction against single photon-loss events (for example, the application of the harmonic-oscillator lowering operator \hat{a}), which is the dominant error mechanism for a cavity functioning as a quantum memory. A photon-loss event on a quantum state $|\psi_L\rangle = \alpha |0_L\rangle + \beta |1_L\rangle$ transforms this state to $|\psi_E\rangle = \hat{a} |\psi_L\rangle = \alpha |E_0\rangle + \beta |E_1\rangle$, with error codewords $|E_0\rangle = |1\rangle$ and $|E_1\rangle = |3\rangle$. The quantum amplitudes α and β are left unchanged despite the loss event. Detecting this single-photon error is straightforward because it results in a photon-number parity flip from even to odd, which is readily measured in our circuit-quantum-electrodynamics system using photon-number parity measurements³⁶. Upon detection of an error event, in principle, a correction unitary operator can be applied that takes $|E_0\rangle \rightarrow |0_L\rangle$ and $|E_1\rangle \rightarrow |1_L\rangle$, preserving the relative quantum amplitudes α and β .

Fock encoding. We also demonstrate the teleported gate using a simple Fock encoding, with basis states $|0\rangle$ and $|1\rangle$, using the lowest two energy levels of the cavity to specify the data qubit. This basis is not a logical encoding according to our definition because it does not allow for quantum error correction; however, by specifying the data qubits in this basis, we can extract an upper bound for the performance of the teleported gate using our current device.

Teleported gate protocol. Before each experimental run, the entire system is initialized in the ground state by an active-feedback cooling sequence. An initial state is encoded onto the data qubits by using the communication qubits as an ancilla for the data qubits. We generate the initial state in the communication qubits and then apply an encoding optimal-control pulse that transfers this state onto the logical basis of the data qubits (Supplementary Information). We design this operation to return the communication qubit back to the ground state so it can be reused for the teleported gate sequence. We then perform the teleported CNOT gate. To analyse the resulting state, we perform tomography on the data qubits using one of two methods: by analysing the logical qubit state or by extracting the Wigner function. We provide a detailed pulse sequence of our experiment and a timing diagram in Extended Data Fig. 3.

In our experiment, it is important to track the reference frame of each data qubit, otherwise the local operations will fail. An important consequence of the Bell-state generation protocol is that the dispersive interaction induces a known, deterministic reference-frame shift on each of the data qubits; we account for this by updating the phase in subsequent steps of the teleported gate protocol. The communication-qubit measurements cause a conditional reference phase shift

on the data qubits dependent on measurement outcome. Tracking these phases accurately is essential for all subsequent operations on the data qubits, and our controller dynamically updates the reference phase of all subsequent operations in real time. We provide further details on the measurement of these data-qubit reference phase shifts in Supplementary Information.

Data-qubit analysis. In the modular architecture, data qubits are designed to be well isolated from the environment and are therefore not measured directly. Instead, we repurpose the communication qubits and use them to measure the state of the data qubits. Using this indirect strategy we perform two types of measurement. In the first type of measurement, we measure the logical (or encoded) state of the data qubits. This is accomplished by first decoding the data-qubit state onto the communication qubit using an optimal-control decoding pulse and then measuring the desired observable on the communication qubit. Because the communication qubit is a transmon, we utilize standard techniques to enact rotations and single-qubit measurements to measure the logical observable. In the second type of measurement, we perform Wigner tomography to fully specify the cavity state³⁶. We perform Wigner tomography using a Ramsey sequence on the communication qubit that maps the photon-number parity of the cavity state onto the state of the communication qubit.

Communication-qubit Bell-state generation. The Bell-state generation occurs while the data qubits store quantum information; the static dispersive interaction between the data and communication qubits, if not accounted for, will naturally entangle the data and communication qubits. Because it is necessary for the two qubits within each module to be disentangled at the end of this step, we modify our Bell-pair generation protocol and implement a refocused RIP sequence²⁸ to echo away this unwanted interaction independently of the data-qubit encoding scheme (Extended Data Fig. 4a).

In our experiment, we utilize a RIP gate of length $T = 300$ ns, with pulse shape $\varepsilon(t) = A[\cos[\pi \cos(\pi t/T)] + 1]$, to minimize the residual photon population left in the bus cavity at the end of the pulse²⁸. We achieve an entangling phase of $\phi_{\text{ent}} = \pi$ in 672 ns and, combined with single communication-qubit rotations, create the Bell state $|\Psi^+\rangle = (|ge\rangle + |eg\rangle)/\sqrt{2}$ with a state fidelity of $97\% \pm 1\%$ (Extended Data Fig. 4b). To perform two-qubit tomography, we choose an over-complete set of 36 single-qubit rotations and use a maximum-likelihood estimate to reconstruct the density operator. We perform 10,000 averages for each tomography setting. Statistical errors are small in this experiment, around 0.2%, as extracted from a bootstrap analysis. Our uncertainty is estimated as the average of several experiments and roughly accounts for the run-to-run variations in our experiment.

Communication-qubit measurement and reset. The success of the teleported CNOT gate requires reliable measurements of each communication qubit. As discussed previously, our JPC-enabled single-qubit readout has assignment fidelities in excess of 99%. In our implementation of the teleported gate, the communication qubits serve dual roles: to store inter-module entanglement and to enable complex data-qubit operations via optimal-control pulses. Therefore, after the measurement of the communication qubits in our protocol, we perform a feedback reset of both communication qubits to the ground state to recycle them for the following single-qubit operations and tomography steps. These measurements are required to be highly quantum-non-demolition to the communication qubit and the data qubits.

We perform the following experiment to test the measurement and the reset. First, we initialize the two communication qubits in an equal superposition of computational states: $|\psi_{\text{init}}\rangle = (|gg\rangle + |ge\rangle + |eg\rangle + |ee\rangle)/2$. Next, we perform measurements on each qubit, allowing the controller to perform real-time state estimation. Conditioned on the measurement results, we apply a π -pulse if the qubit was measured to be in the excited state. Finally, we analyse the state via conditioned state tomography to assess the quality of the reset. The resulting tomograms are shown in Extended Data Fig. 5. We extract state infidelities to the joint ground state $|gg\rangle$ of less than 1% for the case when we measured both qubits in the ground state (outcome '00'). We observe single-qubit infidelities of 2% and 4% when each qubit is measured to be in the excited state. The result from outcome '11' indicates that these infidelities are additive and any crosstalk in the measurement or control is negligible. From these results, we find an average reset infidelity of about 3%, primarily limited by decay during the measurement and subsequent controller latency. From this experiment we establish that our system exhibits highly accurate and quantum-non-demolition single-qubit measurements.

Communication-qubit state tomography. For tomography on the two communication qubits, which are physically transmon qubits, it is convenient to decompose the state in the Pauli basis: $\hat{\rho} = \sum_a p_a \hat{\sigma}_a$, where $\hat{\sigma}_a \in \{\hat{I}, \hat{X}, \hat{Y}, \hat{Z}\}^{\otimes 2}$ are the generalized Pauli operators. We then choose the overcomplete set of single-qubit rotations $\{\hat{I}, \hat{R}_x(\pi), \hat{R}_x(\pm\pi/2), \hat{R}_y(\pm\pi/2)\}^{\otimes 2}$ as tomography operations. Experimentally, we perform independent Z -measurements of each communication qubit, thus extracting two bits of information for each shot. Ideally, this generates the set of computational-state projection operators: $\{\hat{I}_{gg}, \hat{I}_{ge}, \hat{I}_{eg}, \hat{I}_{ee}\}$, where $|\hat{I}_{jk}\rangle = |jk\rangle\langle jk|$. In practice, we calibrate the measurement operators by preparing each of the four computational states and performing our two-bit measurement.

The experimental positive-operator-valued measure elements $\{\hat{P}_{jk}\}$ are then given as $\hat{P}_{jk} = \text{diag}[\text{Pr}('00'|jk), \text{Pr}('01'|jk), \text{Pr}('10'|jk), \text{Pr}('11'|jk)]$; that is, we extract the probability of the four possible measurement outcomes when we prepare $|jk\rangle \in \{|gg\rangle, |ge\rangle, |eg\rangle, |ee\rangle\}$. This analysis assumes that the measurement operator is sensitive to only the \hat{Z} component of the qubit state, and from previous work in which quantum detector tomography was performed³⁷, we find this to be a reasonable assumption.

Data-qubit state tomography. Reliable state tomography is predicated on ensuring small state preparation and measurement errors. However, when considering tomography on the data qubits (and in contrast to tomography on the communication qubits), it is no longer the case that we have a set of trusted operations to effect necessary operations and measurements on these multi-level systems. Therefore, we perform an indirect characterization²⁹ of logical qubit operations \hat{U}_{op} , whereby we perform tomography on the communication qubits for the composite operation $\hat{U}_{\text{dec}}\hat{U}_{\text{op}}\hat{U}_{\text{enc}}$. The protocol begins and ends in the communication-qubit subspace and allows the use of trusted operations and measurements on the communication qubits. Then, by comparing this experiment to the case in which we perform the encoding and decoding pulses, $\hat{U}_{\text{dec}}\hat{U}_{\text{enc}}$, we isolate the performance of only \hat{U}_{op} .

Process tomography. Our approach to performing logical process tomography on the teleported CNOT gate requires performing state tomography on a complete set of two-qubit initial states; here, we choose an overcomplete set of 36 input states $\{|\pm Z_L\rangle, |\pm X_L\rangle, |\pm Y_L\rangle\}^{\otimes 2}$. Experimentally, we apply the appropriate rotation on each communication qubit and use optimal-control pulses to encode the state onto the data qubit. Then, we perform the teleported CNOT gate and subsequently apply a decoding optimal-control pulse to map the data-qubit states onto the communication qubits. With the quantum state contained in the communication qubit, we then perform state tomography on the communication qubits to reconstruct the state. With this set of ideal input states and experimentally reconstructed output states, we perform an inversion to extract the process that maps input states to output states. We represent the reconstructed process using the Pauli transfer matrix $\mathcal{R}_{\text{CNOT}}$, which relates input \mathbf{P}_{in} and output \mathbf{P}_{out} states in the Pauli basis³⁸, $\mathbf{P}_{\text{out}} = \mathcal{R}_{\text{CNOT}}\mathbf{P}_{\text{in}}$. In Extended Data Fig. 6, we present conditioned process tomography results with and without feedforward operations for the binomial encoding; equivalent results for the Fock encoding are provided in Supplementary

Information. We perform a total of six pre- and post-rotations for quantum process tomography (QPT), leading to a total of $6^4 = 1,296$ tomography settings. Each tomography setting consists of 2,500 averages. For each, the statistical error as extracted from a bootstrap analysis is less than 1%; error bars reported in the main text represent an estimate of the run-to-run variation, which is around 2%.

Figures of merit. In this work we use the following two measures for state and process fidelity: (1) the fidelity³⁹ between two states ρ and σ ,

$$\mathcal{F}_{\text{state}}(\rho, \sigma) = \text{tr}(\sqrt{\rho^{1/2}\sigma\rho^{1/2}})^2$$

and (2) the fidelity³⁸ between two processes \mathcal{R}_1 and \mathcal{R}_2 ,

$$\mathcal{F}_{\text{process}}(\mathcal{R}_1, \mathcal{R}_2) = \frac{\text{tr}(\mathcal{R}_1^\dagger \mathcal{R}_2)/d + 1}{d + 1}$$

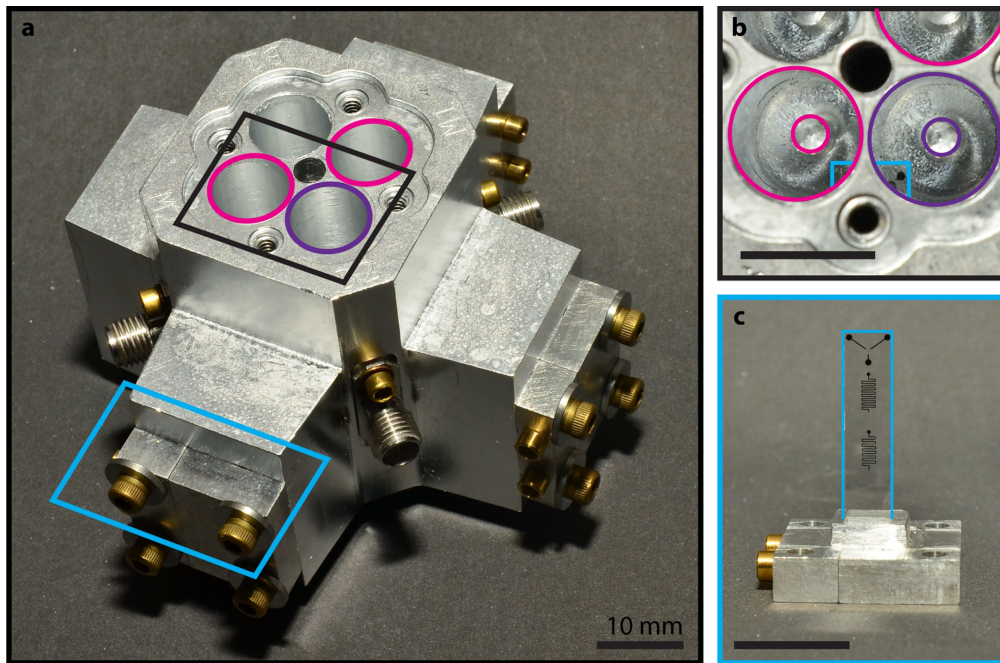
with $d = 2n$ and n the number of qubits. We use the standard formula⁴⁰ to calculate the concurrence \mathcal{C} .

The process fidelity calculated above is similar to the average gate fidelity, which for two processes \mathcal{E}_1 and \mathcal{E}_2 is generally defined as

$$\mathcal{F}_{\text{avg}} \equiv \int \mathcal{F}_{\text{state}}[\mathcal{E}_1(\rho), \mathcal{E}_2(\rho)] d\psi$$

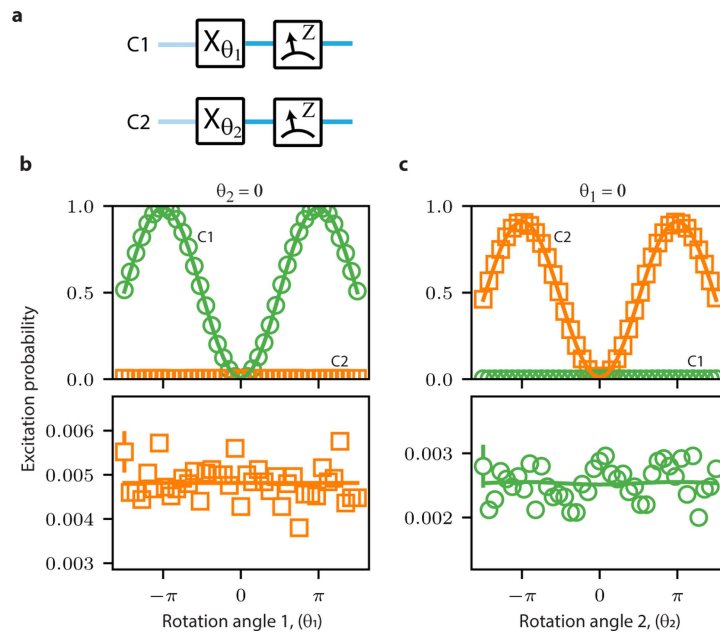
Data availability. The data that support the findings of this study are available from the corresponding authors on reasonable request.

36. Vlastakis, B. et al. Deterministically encoding quantum information using 100-photon Schrödinger cat states. *Science* **342**, 607–610 (2013).
37. Blumoff, J. Z. et al. Implementing and characterizing precise multiqubit measurements. *Phys. Rev. X* **6**, 031041 (2016).
38. Chow, J. M. et al. Universal quantum gate set approaching fault-tolerant thresholds with superconducting qubits. *Phys. Rev. Lett.* **109**, 060501 (2012).
39. Gilchrist, A., Langford, N. K. & Nielsen, M. A. Distance measures to compare real and ideal quantum processes. *Phys. Rev. A* **71**, 062310 (2005).
40. Wootters, W. K. Entanglement of formation of an arbitrary state of two qubits. *Phys. Rev. Lett.* **80**, 2245–2248 (1998).



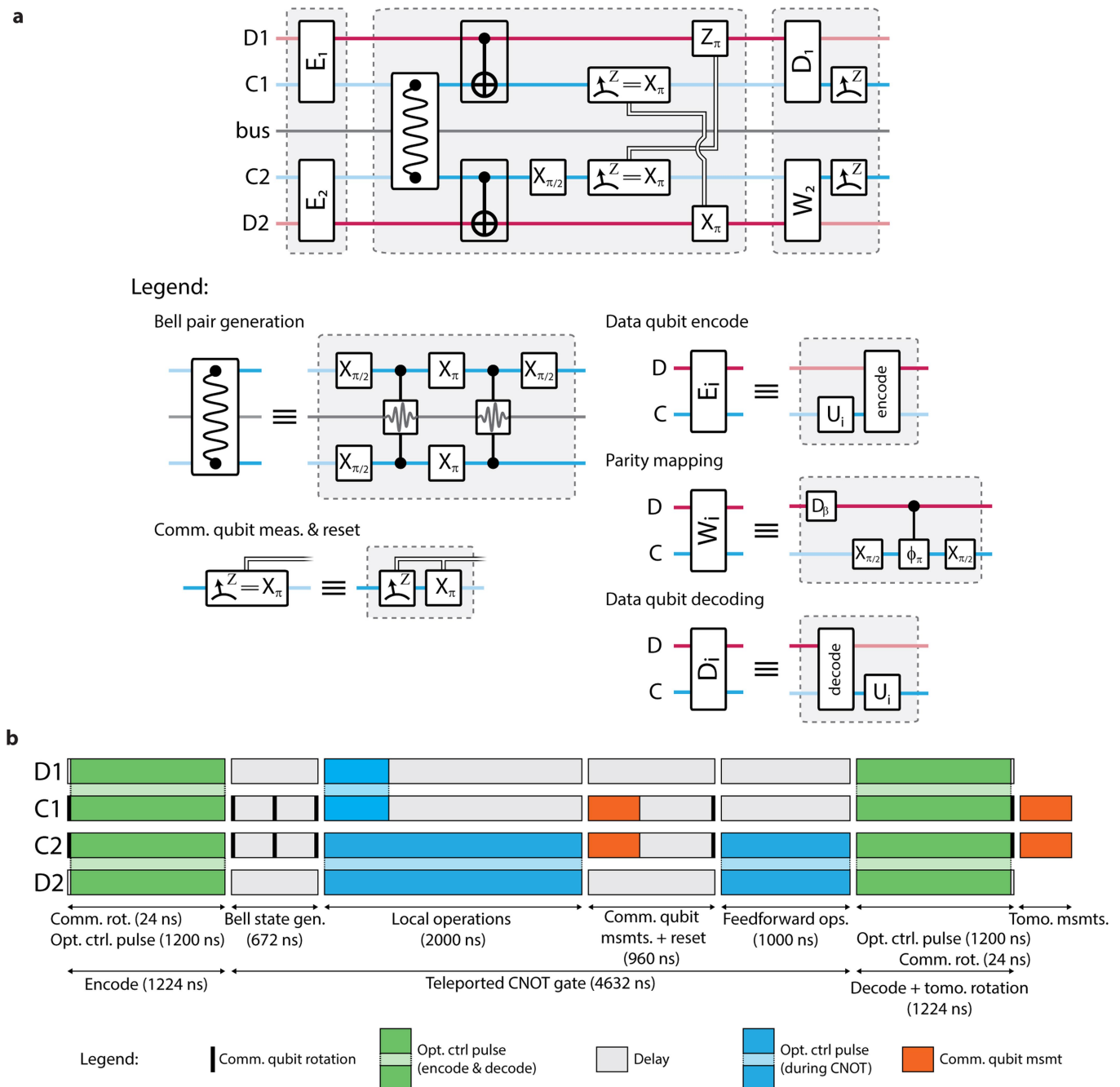
Extended Data Fig. 1 | Overview of the physical device. **a**, Photograph of the full device assembly. The main body of the device is constructed from high-purity (99.99%) aluminium and contains four coaxial $\lambda/4$ three-dimensional cavities, three of which are used. The cavities that serve as the data qubits and the bus are outlined in magenta and purple, respectively. A detailed photograph of the cavities is shown in **b**. Two clamps anchor each sapphire chip; one is highlighted in cyan and detailed in **c**. The visible connectors are input ports for each cavity; the input and output ports for the transmon and readout resonators are on the underside of the device

and therefore not visible. **b**, Top-down photograph of the cavities. We illustrate the three cavities using the same colour scheme as in **a**; the inner circle represents the inner conductor that defines the cavity mode; the cyan outline shows the sapphire chip inserted into the device package. Also visible are the antenna pads of the transmon that enable coupling to each cavity. **c**, Photograph of the sapphire chip on which the transmon and readout resonators are fabricated. The sapphire chip is outlined in cyan and contains several elements: from the top of the figure moving down, the Y-shaped transmon qubit, the readout resonator and the Purcell filter.



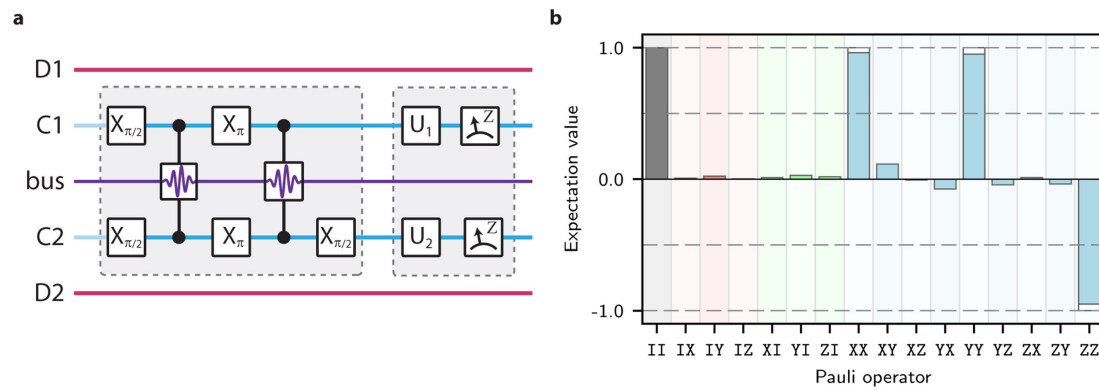
Extended Data Fig. 2 | Assessing the independence of communication-qubit measurements. **a**, Rabi experiment pulse sequence to extract measurement crosstalk. After initializing both communication qubits in the ground state, both qubits are rotated by \hat{X} rotations, with independent angles θ_1 and θ_2 for C1 and C2, respectively. Subsequently, measurements are performed on modules 1 and 2 and the result is recorded. **b**, **c**, Measurement crosstalk experimental results. For **b** (**c**), C2 (C1) is kept in the ground state, and a Rabi experiment is performed on C1 (C2). The measurement results are shown for C1 (green circles) and C2 (orange squares). For clarity, we describe the results focusing on **b**; the discussion is the same for **c**, save for swapping C1 and C2. Top, the C1 measurement results illustrate high-contrast oscillations, whereas the C2 measurement

results remain close to zero, as expected when the communication-qubit measurements are independent. Bottom, close-up for measurement results on C2. The lack of structure in the data indicates that the measurement of C2 does not infer any information about the state of C1. To estimate the measurement crosstalk, we perform sinusoidal fits to the data by fixing the frequency and phase of the oscillation and extracting an amplitude and offset. Each data point in this experiment corresponds to 25,000 experiments. For data in the top panels, error bars are much smaller than the marker; for data in the bottom panels, we represent a typical error bar to be within the spread of the points. The slightly reduced contrast in **c** is specific to this calibration experiment, and potentially due to drifts in the transmon relaxation rate during the many hours of acquisition.



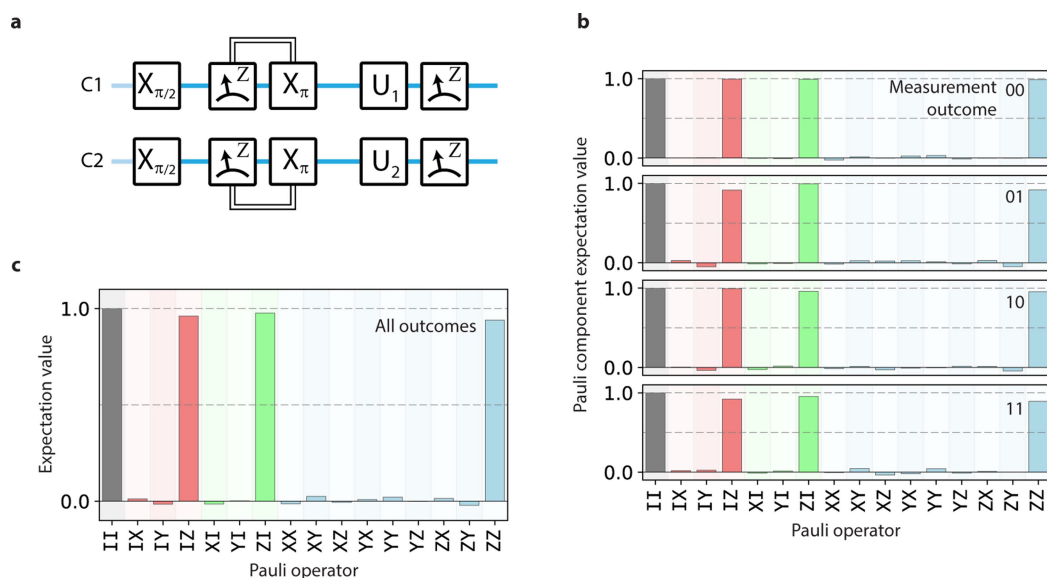
Extended Data Fig. 3 | Implementation of the teleported CNOT gate. **a**, Detailed circuit diagram for the teleported CNOT gate protocol. Top, pulse sequence for an example experiment. Bottom, legend for specific circuit blocks. In the first panel, we show our sequence for encoding quantum information onto the data qubit. In the second panel, we illustrate our implementation of the teleported CNOT gate. We show the pulse sequence used to generate the communication-qubit Bell state. For the communication-qubit measurements, we apply a $\pi/2$ rotation on C2 to measure \hat{X} . After the measurement we also perform a measurement-based reset of C1 and C2 before performing feedforward operations on the data qubits. In the third panel, we detail two possible sequences for extracting the data-qubit state. For module 1, we perform logical tomography on the data qubits by decoding the data qubit onto the communication qubit and

performing the appropriate tomography rotations on the communication qubit. For module 2, we perform Wigner tomography by performing a parity-mapping sequence on the communication qubit. **b**, Teleported CNOT gate timing diagram. The teleported CNOT gate is illustrated taking the relative timing of each element into account. The diagram is colour-coded with the following designations: black, single communication-qubit rotations; green, encode and decode (optimal control) operations; blue, teleported CNOT gate local operations (also optimal control); orange, measurements. This presentation provides a visual representation of the relative durations of each part of the protocol. Our implementation of the teleported CNOT gate takes a total of approximately 4.6 μ s.



Extended Data Fig. 4 | Communication-qubit Bell state. a, Pulse sequence for generating the communication-qubit Bell pair. After generating the Bell state (first block), quantum state tomography is performed on both of the qubits to assess the quality of the entangled state.

b, Characterizing the communication-qubit Bell pair. Experimentally measured Pauli vector components of the two communication qubits are shown. The generated state is $(|ge\rangle + |eg\rangle)/2$, with the ideal values denoted as hollow bars.

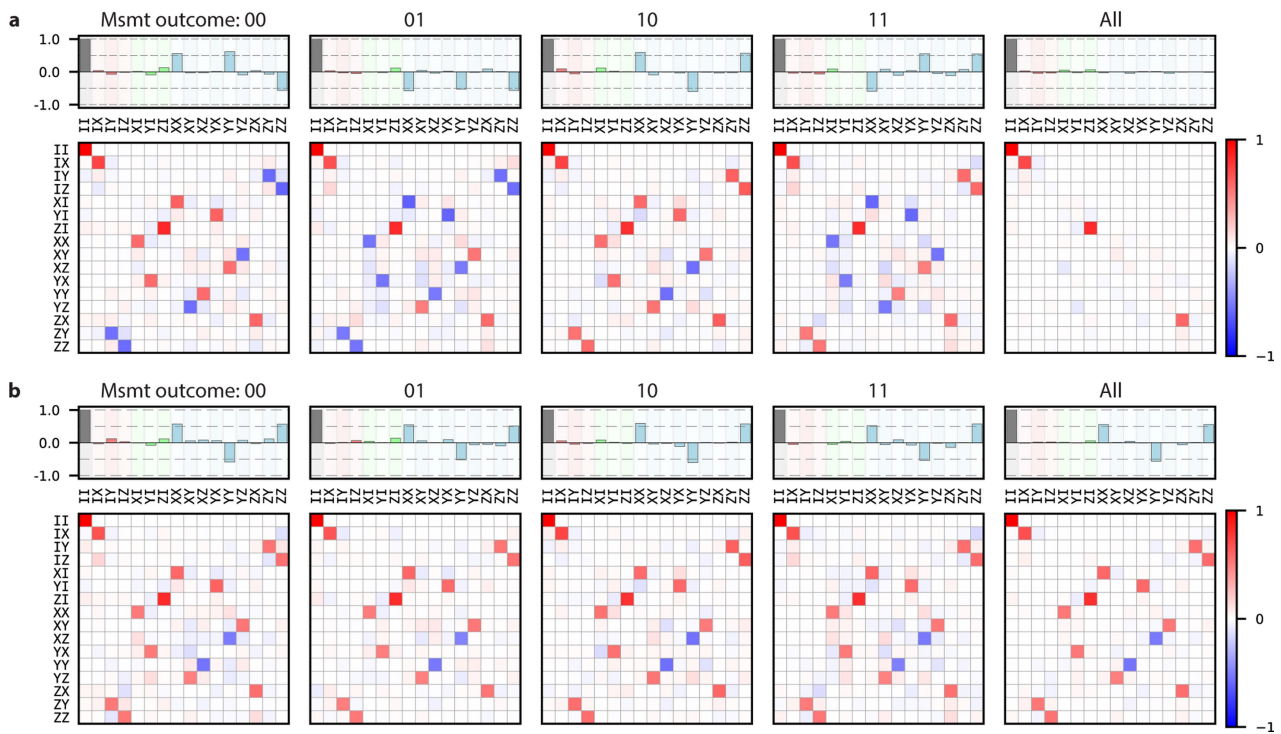


Extended Data Fig. 5 | Communication-qubit measurement and reset.

a, Pulse sequence for testing communication-qubit measurement and reset. The two communication qubits (transmons) are initialized in the joint state $(|gg\rangle + |ge\rangle + |eg\rangle + |ee\rangle)/2$. The two qubits are then measured and if the measurement indicates that the state is projected to $|e\rangle$ a π -pulse is applied to flip the state to the ground state. Conditional quantum state tomography is performed to analyse the quality of measurement and reset. This measurement and reset protocol is used in the teleported gate.

b, Experimentally measured Pauli vector components conditioned on the measurement outcome. We assign a '0' ('1') to indicate that the measurement projected the qubit to be in $|g\rangle$ ($|e\rangle$). For all outcomes, we

find high fidelity to the two-qubit ground state $|gg\rangle$, as expected, with ground-state fidelities of {00, 99.3%; 01, 95.7%; 10, 97.7%; 11, 94.2%}. From these results, we establish that the measurement and feedback processes for each qubit are independent; from the single-qubit reset infidelities, we expect a measurement fidelity of $1 - (0.993 - 0.957) - (0.993 - 0.977) = 0.948$, which is consistent with the result for measurement outcome 11. **c**, Experimentally measured state after measurement-based reset. Measurement results from **b** are combined, and the compiled results illustrate that the reset protocol is high-fidelity and independent of the measurement outcome. The fidelity of this reconstructed two-qubit state to $|gg\rangle$ is 96.9%.



Extended Data Fig. 6 | Extended binomial QPT data. For each panel, we plot the process matrix in the Pauli transfer representation (bottom) and a reconstructed state represented in the Pauli basis (top). For the reconstructed state, we choose the input state $(|0\rangle + |1\rangle)|0\rangle/\sqrt{2}$, which should result in the Bell state $|\Phi^+\rangle = (|00\rangle + |11\rangle)/\sqrt{2}$ when the CNOT gate is applied. The ideal process for each panel is represented by the dominant components taken to ± 1 and small components taken to 0. **a**, Conditioned QPT results when the feedforward operations are not applied. The first four panels (labelled '00', '01', '10' and '11') represent the processes conditioned on measurement outcome. Each has qualitatively the same features (for example, the same non-zero elements of the process matrix); however, the differing signs between the four outcomes indicate that each process is modified by single-qubit operations. When all

measurement results are combined (labelled 'All'), most of the features are washed away and only certain Pauli operators are left invariant by the process: $\{II, IX, ZI, ZX\}$. These operators are exactly the feedforward operations that would normally be applied. This behaviour can also be observed in the state tomography results (top), in which each measurement outcome heralds a different Bell state ($|\Psi^+\rangle, |\Psi^-\rangle, |\Phi^+\rangle, |\Phi^-\rangle$); when taken all together, the states add incoherently, resulting in a completely mixed state. **b**, Conditioned QPT results when the feedforward operations are applied. Here, all measurement outcomes (00, 01, 10, 11) indicate the same process, that of the CNOT process. Therefore, when the measurement outcomes are all taken together (All), the compiled process is that of a CNOT gate.

Absolute timing of the photoelectric effect

M. Ossiander^{1,2,5*}, J. Riemensberger^{1,2,5}, S. Nepp³, M. Mittermair¹, M. Schäffer^{1,2}, A. Duensing¹, M. S. Wagner¹, R. Heider¹, M. Wurzer¹, M. Gerl^{1,2}, M. Schnitzenbaumer¹, J. V. Barth¹, F. Libisch⁴, C. Lemell⁴, J. Burgdörfer⁴, P. Feulner¹ & R. Kienberger^{1,2*}

Photoemission spectroscopy is central to understanding the inner workings of condensed matter, from simple metals and semiconductors to complex materials such as Mott insulators and superconductors¹. Most state-of-the-art knowledge about such solids stems from spectroscopic investigations, and use of subfemtosecond light pulses can provide a time-domain perspective. For example, attosecond (10^{-18} seconds) metrology allows electron wave packet creation, transport and scattering to be followed on atomic length scales and on attosecond timescales^{2–7}. However, previous studies could not disclose the duration of these processes, because the arrival time of the photons was not known with attosecond precision. Here we show that this main source of ambiguity can be overcome by introducing the atomic chronoscope method, which references all measured timings to the moment of light-pulse arrival and therefore provides absolute timing of the processes under scrutiny. Our proof-of-principle experiment reveals that photoemission from the tungsten conduction band can proceed faster than previously anticipated. By contrast, the duration of electron emanation from core states is correctly described by semiclassical modelling. These findings highlight the necessity of treating the origin, initial excitation and transport of electrons in advanced modelling of the attosecond response of solids, and our absolute data provide a benchmark. Starting from a robustly characterized surface, we then extend attosecond spectroscopy towards isolating the emission properties of atomic adsorbates on surfaces and demonstrate that these act as photoemitters with instantaneous response. We also find that the tungsten core-electron timing remains unchanged by the adsorption of less than one monolayer of dielectric atoms, providing a starting point for the exploration of excitation and charge migration in technologically and biologically relevant adsorbate systems.

Two complementary methods of attosecond metrology, namely, the attosecond streak camera⁸ and RABITT interferometry⁹, are capable of providing attosecond-scale timing information on electron emission. The ultimate aim—to determine the absolute duration of the photoelectric effect from a solid, that is, the temporal sequence of events between the arrival time of an ionizing extreme-ultraviolet (XUV) photon at its surface and photoelectron emission into vacuum—has remained a major challenge. Here we report an approach that can overcome this difficulty: by using adsorbed atoms as a chronoscope whose absolute photoionization timing can be determined in concurrent gas-phase measurements, the absolute timing of photoemission from solid surfaces can now be clocked with attosecond precision. As proof of the concept, we examine the photoelectric effect for the dense-packed W(110) surface of a tungsten crystal at 105 eV photon energy using an attosecond streak camera. The measurement geometry is depicted in Fig. 1a. We adsorb iodine atoms on a metallic surface: these atoms serve as chronoscope species. When the XUV light pulse arrives at the surface, it stimulates electron expulsion from these clock atoms. The appearance of these electrons in the ionization continuum marks the light pulse's arrival time at the surface after correction for the atomic photoemission delay of the adatom^{10,11}. The light pulse

simultaneously propagates into the crystal and photoexcites electrons from both localized core orbitals and delocalized conduction bands of W(110). Thus, Bloch wave packets in high-lying conduction bands evolve, and some travel towards the surface and exit the crystal. Using a defined near-infrared (NIR) electric waveform (temporal full-width at half-maximum $\tau_{\text{FWHM}} \approx 4$ fs, carrier wavelength $\lambda_{\text{carrier}} = 780$ nm), we encode the instant of an electron's appearance in the ionization continuum as an observable momentum modulation ('streaking'⁸). The NIR light is reflected and refracted at the metal surface such that the electric field components causing streaking are suppressed within the crystal³. Therefore, electrons originating from the crystal interior are only momentum-modulated once they have left the solid, and their sojourn time in the crystal can be retraced. The ångström-scale proximity of the chronoscope atoms to the surface eliminates uncertainties in the synchronization of the two light pulses due to Gouy phase differences, propagation and modification of the streaking field by reflection off the surface¹². The last resulted in uncertainties of several tens of attoseconds in previous timing measurements of the photoelectron creation process^{13,14}. We then gauge the emission delay of our clock atoms to achieve a true absolute delay timing. This is accomplished in a gas-phase streaking experiment using a mixture of chronoscope species and helium, illustrated in Fig. 1b. For the latter, the absolute photoemission delay is well tested and exactly computed by theory^{15–17}. The complete experimental timing sequence is depicted in Fig. 1c.

Owing to its spectral isolation from helium and tungsten surface photoelectrons, we employ core-level photoemission from the 4*d* inner-shell orbitals of iodine (I4*d*) as a time-zero marker (see Fig. 2a, b). Core-level emission delays are largely independent of the chemical environment, as has been shown to hold for crystals⁵. The large photoabsorption cross-section of the I4*d* channel enhanced by the so-called giant dipole resonance allows for sufficient signal strength at a wide range of iodine coverage. Hence, we can quantify and eliminate residual transport and screening effects of the iodine overlayer by systematic variation of the adatom density and extrapolation to zero coverage. To obtain the moment of photon arrival at the surface from the chronoscope photoemission, the absolute I4*d* emission timing was determined in gauge measurements conducted on a gaseous mixture of small iodine-containing molecules and helium. Further considerations on the choice of chronoscope species, the gauge measurements, the comparability of molecular and adsorbate photoemission and the delay extraction method are detailed in Methods. The experimental setup is described elsewhere^{18,19}.

Experimental results are summarized in Figs. 1c and 2c. In the gas-phase gauge measurements, we find a relative delay of $\Delta\tau_{\text{I4d-He1s}} = (31 \pm 3)$ as between the escape of I4*d* and helium ground state (He1s) electrons. The absolute photoemission delay of photoelectrons escaping the 1s orbital in helium at 105 eV photon energy is^{11,15} $\tau_{\text{He1s}} = -5.0$ as. This yields an absolute I4*d* delay of $\tau_{\text{I4d}} = \Delta\tau_{\text{I4d-He1s}} + \tau_{\text{He1s}} = (26 \pm 3)$ as. All reported uncertainties represent 95% confidence. The delayed emission of I4*d* compared to He1s is mainly caused by an increased Eisenbud–Wigner–Smith (EWS) delay^{20–22} due to the giant dipole resonance²³ (see also Methods).

¹Physik-Department, Technische Universität München, Garching, Germany. ²Max-Planck-Institut für Quantenoptik, Garching, Germany. ³Helmholtz-Zentrum Berlin für Materialien und Energie, Berlin, Germany. ⁴Institute for Theoretical Physics, Vienna University of Technology, Vienna, Austria. ⁵These authors contributed equally: M. Ossiander, J. Riemensberger. *e-mail: marcus.ossiander@mpq.mpg.de; reinhard.kienberger@tum.de

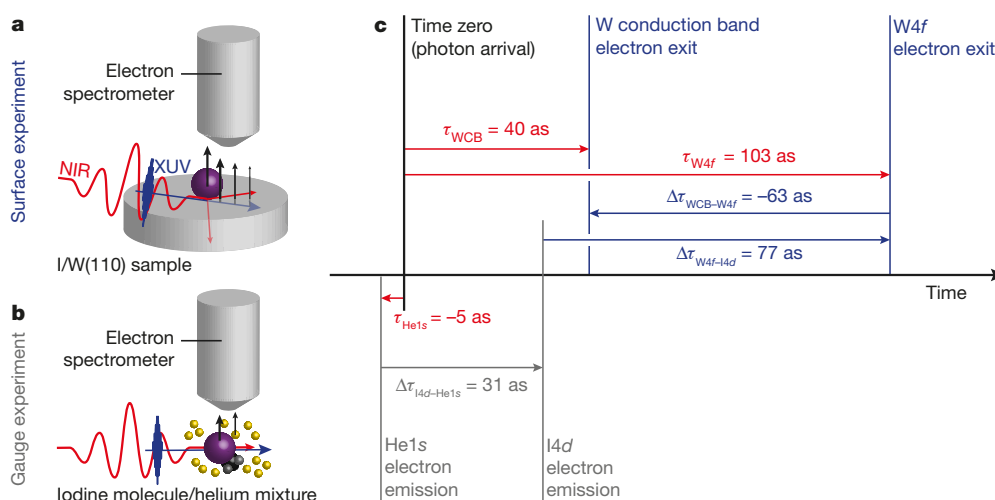


Fig. 1 | The atomic chronoscope method. **a**, Surface experiment. An XUV (blue) light pulse launches photoelectrons (black arrows) from a W(110) surface (grey) and from an iodine chronoscope (purple) on top. An NIR (red) laser pulse encodes the appearance time of photoelectrons above the crystal as a momentum shift, which is resolved using an

electron spectrometer. **b**, Gas-phase gauge experiment. The delay between photon absorption by the chronoscope and photoelectron appearance is determined by comparing to helium (yellow). **c**, Full timing sequence. Absolute delays are depicted in red, surface-experiment delays in blue and gauge-measurement delays in grey.

With the absolute time reference at hand, we can precisely determine the absolute electron exit delay for tungsten 4f (W4f) core electrons escaping the W(110) surface. A linear fit to the data in Fig. 2c reveals a coverage-dependent exit delay of $\tau_{\text{W4f}}(\theta) = \Delta\tau_{\text{W4f-I4d}}(\theta) + \tau_{\text{I4d}} = (77 \pm 5) \text{ as} + \theta(8 \pm 7) \text{ as} + \tau_{\text{I4d}}$ for W4f electrons as a function of the fractional iodine surface coverage θ in units of saturated monolayers (sat. ML). Importantly, the small influence of the iodine adlayer on the delay over the wide range of examined coverages allows assessment of the behaviour of pristine W(110) via extrapolation, which yields an absolute W4f exit delay $\tau_{\text{W4f}}^{\text{clean}} = (103 \pm 6) \text{ as}$ for a bare surface.

By contrast, the average valence-electron timing of iodine-covered W(110) (denoted I/W(110)) features appreciable coverage dependence due to spectral overlap of weakly bound electrons from iodine and the tungsten conduction band. We thus performed additional measurements on a pristine W(110) surface, and found a W4f to tungsten conduction-band (WCB) delay $\tau_{\text{W4f-WCB}}^{\text{clean}} = (63 \pm 6) \text{ as}$. Owing to the high XUV photon energy, we can observe emission from the full Brillouin zone without delay artefacts due to misalignment of the electron detection direction and streaking laser field polarization²⁴. By referencing to the absolute W4f exit delay, we determine an absolute tungsten

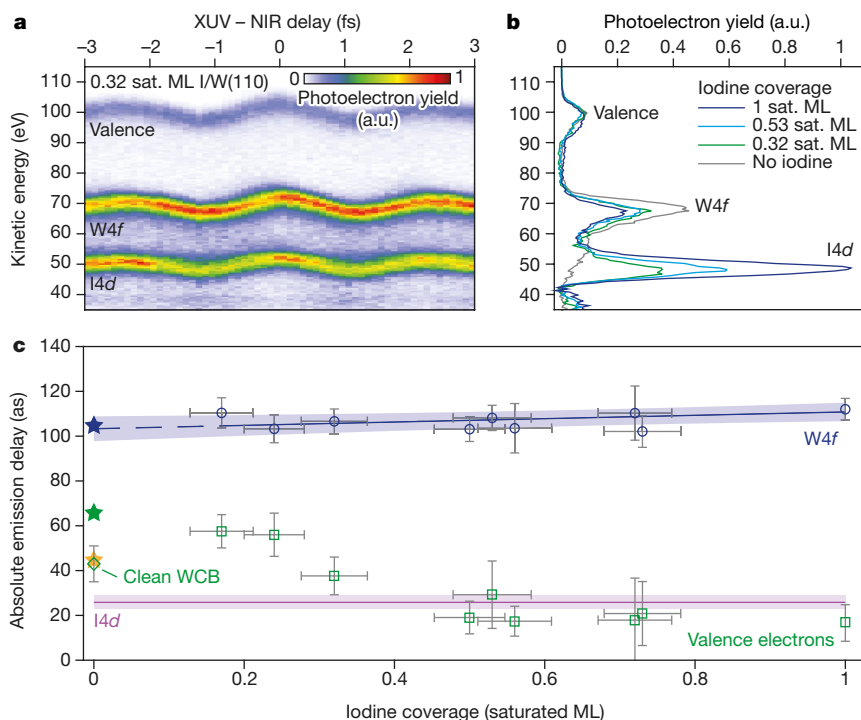


Fig. 2 | Absolute timing of W(110) photoemission at 105 eV photon energy. **a**, Representative I/W(110) streaking spectrogram. Delays are encoded as phase shifts of the kinetic-energy oscillations. **b**, XUV-only photoemission spectra for I/W(110) and pristine W(110). **c**, Iodine-coverage-dependent photoemission timing. Shown are W4f exit delay averages at coinciding iodine coverage (blue circles), the extrapolation to

the pristine surface (blue line), I/W(110) valence-electron delays (green squares), the pristine W(110) conduction-band timing (green diamond), the I4d reference (purple) and transport simulation results (stars, W4f (blue), WCB with (yellow) and without (green) surface state influence). Errors represent 95% confidence intervals. Vertical error bars are calculated assuming a Student's *t*-distribution.

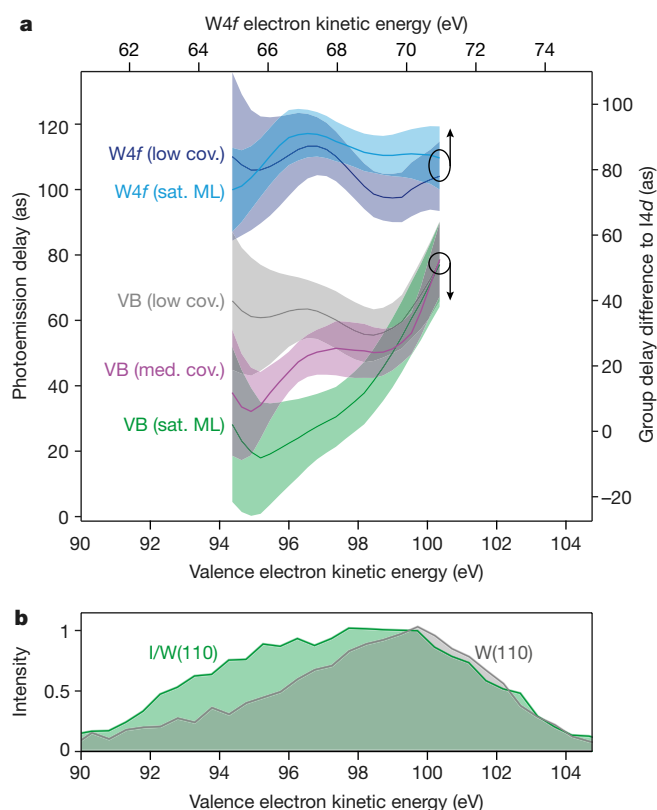


Fig. 3 | Timing of adsorbate photoemission. **a**, Energy-resolved emission delay of W4f core and I/W(110) valence-band (VB) photoelectrons for different iodine coverages. Core-level emission timing (blue lines) is unaffected by the iodine coverage (cov.) within the 95% confidence interval (shaded areas). The emission delay of valence states is substantially reduced, and its dispersion increases by adsorbing iodine (grey, purple and green lines). **b**, Photoelectron spectra of an iodine-saturated (green) and a clean (grey) W(110) surface. Iodine mainly contributes to the lower-kinetic-energy part of the valence-photoelectron emission, visible in the spectra and in the energy-resolved photoemission delays.

conduction-band exit delay $\tau_{\text{WCB}}^{\text{clean}} = \tau_{\text{W4f}}^{\text{clean}} - \Delta\tau_{\text{W4f-WCB}}^{\text{clean}} = (40 \pm 9)$ as. The small conduction-band photoemission delay for clean tungsten is surprising, as it is comparable to the EWS delay in gas-phase I4d photoemission.

A major advantage of absolute emission timing is that the duration of the creation process for each observed photoemission feature is individually recorded. Thus, all delay contributions, even those cancelling in relative measurements²⁵, are uncovered by the present absolute measurement, enhancing the importance of the extracted timing information for benchmarking theoretical models and allowing their direct interpretation. Tungsten inner-shell photoemission is reproduced quantitatively by transport simulations based on a three-step model (see Methods): the primary photoexcitation of a W4f electron contributes an EWS delay of about 10 as (see ref. ⁵ and its Supplementary Information). Subsequently, the electron propagates to the crystal surface and electron streaking starts at the jellium edge³, located about 1.1 Å above the topmost tungsten layer. The average escape depth corresponds to the mean free path for inelastic scattering²⁶, approximately 4.1 Å. Adding the transport time for free-electron propagation to the EWS delay predicts a total exit delay of $\tau_{\text{W4f}} = 85$ as, slightly less than measured in our experiment. Including inelastically scattered conduction-band electrons with an energy close to the W4f line yields 105 as, in almost perfect agreement with the measured data. This is markedly different from ref. ², where a reduced group velocity for W4f electrons was proposed as a reason for the observed relative core-level to conduction-band delay for comparable emission angle integration at slightly lower photon energy.

Electron ejection from the conduction band proceeds substantially faster. Using the effective emission depth of photoionized conduction-band electrons²⁶, about 4.2 Å, predicts an exit delay of $\tau_{\text{WCB}} = 66$ as for conduction-band electrons, considerably larger than measured. As the relevant band structure does not feature group velocities much larger than that of a free electron in the relevant energy region²⁷, we attribute the remaining discrepancy to a major contribution of surface states to the tungsten conduction-band emission. Surface states of W(110) have been studied theoretically²⁸ and experimentally²⁹. Using density-functional-theory (DFT) calculations we identify surface states located in the surface bandgap of the clean substrate (see Methods). From their depth distribution, we determine an exit delay of $\tau_{\text{SS}} \approx 11$ as. Including a surface state contribution with a spectral weight of 1/3 relative to the total conduction-band emission^{29,30} in the transport simulation reduces the average conduction-band exit delay to $\tau_{\text{CB}} \approx 45$ as and reproduces the experimental value for the pristine W(110) surface. In turn, quenching of surface states by adsorbates covering the surface can cause a larger exit delay and the non-monotonic behaviour of τ_{CB} at low coverages (Fig. 2c). These findings highlight that proper accounting for the initial creation, origin, transport and scattering of electrons is imperative for the proper description of the photoelectric effect.

To scrutinize the electronic dynamics of the adsorbate surface layer, we resolve the energy of the exit delays using a generalized-projection algorithm^{15,31}. Results are presented in Fig. 3a. Spectrally averaged delays match those obtained from the original retrieval method. The valence photoelectron spectrum in Fig. 3b is dominated by emission from the tungsten conduction band in the high-kinetic-energy region and by adsorbate iodine valence states in the low-kinetic-energy flank, consistent with DFT calculations (see Methods). Whereas the tungsten conduction-band (high-energy region) delay is only marginally altered by the adsorption of chronoscope atoms, the adsorbate (low-energy flank) delay decreases with increasing iodine surface coverage. Comparing the photoemission delay for different adsorbate densities thus directly reflects the shift of the valence photoelectron origin from within the first crystal layers to the iodine above the surface. Accounting for residual contributions of the tungsten conduction band reveals an iodine valence (IV) photoemission delay of $\tau_{\text{IV}} = (8 \pm 19)$ as. Atomic adsorbates therefore enable the creation of photoelectrons with negligible lag in time even from solid systems. The stability of the W4f timing proves that crystal core-state photoemission is robust against the addition of non-screening adsorbates. Hence, future studies can directly explore the time domain evolution of complex adsorbates by using the present surface core-electron results as a reference.

The atomic chronoscope is a viable tool with which one can determine the absolute timing sequence of the photoelectric effect in condensed matter systems. All terms contributing to its duration are captured, allowing the analysis of future measurements without ambiguity, even those performed on complex interfacial architectures. Clocking a photoelectron's creation is a direct gateway to its phase; hence our results complement and challenge our understanding of the electronic structure of solids, which has so far been shaped by spectral investigations. Surface-adsorbate systems could allow the creation of nanoscopic switches for molecular electronics, the improvement of the efficiency of chemical reactions using heterogeneous catalysis, or the inexpensive harvesting of solar energy. The initial dynamics that lead to photon absorption and subsequent chemical response are the key to systematic design of devices. The photoemission delay of adsorbates probes the local environment around the adsorption sites³². Delay measurements using the present technique and timing reference could thus serve as attosecond probes and help to unravel the non-equilibrium dynamics that initiate photochemistry.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0503-6>.

Received: 12 April 2018; Accepted: 26 July 2018;
Published online 19 September 2018.

1. Damascelli, A. Probing the electronic structure of complex systems by ARPES. *Phys. Scr. T* **109**, 61–62 (2004).
2. Cavaliere, A. L. et al. Attosecond spectroscopy in condensed matter. *Nature* **449**, 1029–1032 (2007).
3. Neppl, S. et al. Direct observation of electron propagation and dielectric screening on the atomic length scale. *Nature* **517**, 342–346 (2015).
4. Seiffert, L. et al. Attosecond chronoscopy of electron scattering in dielectric nanoparticles. *Nat. Phys.* **13**, 766–770 (2017).
5. Siek, F. et al. Angular momentum-induced delays in solid-state photoemission enhanced by intra-atomic interactions. *Science* **357**, 1274–1277 (2017).
6. Chen, C. et al. Distinguishing attosecond electron–electron scattering and screening in transition metals. *Proc. Natl Acad. Sci. USA* **114**, E5300–E5307 (2017).
7. Tao, Z. et al. Direct time-domain observation of attosecond final-state lifetimes in photoemission from solids. *Science* **353**, 62–67 (2016).
8. Kienberger, R. et al. Atomic transient recorder. *Nature* **427**, 817–821 (2004).
9. Müller, H. G. Reconstruction of attosecond harmonic beating by interference of two-photon transitions. *Appl. Phys. B* **74**, s17–s21 (2002).
10. Schultze, M. et al. Delay in photoemission. *Science* **328**, 1658–1662 (2010).
11. Pazourek, R., Nagele, S. & Burgdörfer, J. Attosecond chronoscopy of photoemission. *Rev. Mod. Phys.* **87**, 765–802 (2015).
12. Lucchini, M. et al. Light-matter interaction at surfaces in the spatiotemporal limit of macroscopic models. *Phys. Rev. Lett.* **115**, 137401 (2015).
13. Locher, R. et al. Energy-dependent photoemission delays from noble metal surfaces by attosecond interferometry. *Optica* **2**, 405–410 (2015).
14. Kasmi, L. et al. Effective mass effect in attosecond electron transport. *Optica* **4**, 1492–1497 (2017).
15. Ossiander, M. et al. Attosecond correlation dynamics. *Nat. Phys.* **13**, 280–285 (2017).
16. Palacios, A., McCurdy, C. W. & Rescigno, T. N. Extracting amplitudes for single and double ionization from a time-dependent wave packet. *Phys. Rev. A* **76**, 043420 (2007).
17. Pazourek, R., Feist, J., Nagele, S. & Burgdörfer, J. Attosecond streaking of correlated two-electron transitions in helium. *Phys. Rev. Lett.* **108**, 163001 (2012).
18. Magerl, E. et al. A flexible apparatus for attosecond photoelectron spectroscopy of solids and surfaces. *Rev. Sci. Instrum.* **82**, 063104 (2011).
19. Cavaliere, A. L. et al. Intense 1.5-cycle near infrared laser waveforms and their use for the generation of ultra-broadband soft-X-ray harmonic continua. *New J. Phys.* **9**, 242 (2007).
20. Eisenbud, L. *The Formal Properties of Nuclear Collisions*. PhD thesis, Princeton Univ. (1948).
21. Wigner, E. P. Lower limit for the energy derivative of the scattering phase shift. *Phys. Rev.* **98**, 145–147 (1955).
22. Smith, F. T. Lifetime matrix in collision theory. *Phys. Rev.* **118**, 349–356 (1960).
23. Huppert, M., Jordan, I., Baykusheva, D., von Conta, A. & Wörner, H. J. Attosecond delays in molecular photoionization. *Phys. Rev. Lett.* **117**, 093001 (2016).
24. Heuser, S. et al. Angular dependence of photoemission time delay in helium. *Phys. Rev. A* **94**, 063409 (2016).
25. Neppl, S. et al. Attosecond time-resolved photoemission from core and valence states of magnesium. *Phys. Rev. Lett.* **109**, 087401 (2012).
26. Tanuma, S., Powell, C. J. & Penn, D. R. Calculations of electron inelastic mean free paths. II. Data for 27 elements over the 50–2000 eV range. *Surf. Interface Anal.* **17**, 911–926 (1991).
27. Krasovskii, E. E. Attosecond spectroscopy of solids: streaking phase shift due to lattice scattering. *Phys. Rev. B* **84**, 195106 (2011).
28. Mirhosseini, H., Flieger, M. & Henk, J. Dirac-cone-like surface state in W(110): dispersion, spin texture and photoemission from first principles. *New J. Phys.* **15**, 033019 (2013).
29. Pi, T.-W., Hong, L.-H. & Cheng, C.-P. Synchrotron-radiation photoemission study of Ba on W(110). *Phys. Rev. B* **58**, 4149–4155 (1998).
30. Riemensberger, J. *Time-Frequency-Resolved Absolute Time Delay of the Photoelectric Effect*. PhD Thesis, Technische Universität München (2018).
31. Yakovlev, V. S., Gagnon, J., Karpowicz, N. & Krausz, F. Attosecond streaking enables the measurement of quantum phase. *Phys. Rev. Lett.* **105**, 073001 (2010).
32. Kazansky, A. K. & Echenique, P. M. Theoretical study of the ionization of an alkali atom adsorbed on a metal surface by a laser-assisted subfemtosecond pulse. *Phys. Rev. B* **81**, 075440 (2010).

Acknowledgements We acknowledge discussions with M. Schultze, experimental support by A. Kim and A. Schiffrin and infrastructural support by F. Krausz. This work was supported by the Max Planck Society, the Deutsche Forschungsgemeinschaft Cluster of Excellence, Munich Centre for Advanced Photonics, a Consolidator Grant from the European Research Council (ERC-2014-CoG AEDMOS), LASERLAB-EUROPE (grant agreement number 654148, European Union's Horizon 2020 research and innovation programme), FWF Austria (SFB-041 ViCoM, SFB-049 NextLite) and COST Action CM1204 (XLIC). Calculations were performed using the Vienna Scientific Cluster (VSC).

Reviewer information Nature thanks M. Chini and T. Fennel for their contribution to the peer review of this work.

Author contributions M.O., J.R., S.N., M.M., M. Schäffer, A.D., M.S.W., R.H., M.W., M.G. and M. Schnitzenbaumer carried out the experiments. M.O. and J.R. analysed the experimental data. F.L., C.L. and J.B. performed the electron transport and DFT calculations. M.O. wrote the initial manuscript. J.V.B., P.F. and R.K. supervised the study. All authors discussed and reviewed the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0503-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.O. or R.K.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample preparation and coverage calibration. Samples were prepared in situ in $\sim 10^{-10}$ mbar ultrahigh vacuum. Commercially available polished W crystals with (110) oriented surfaces were cleaned via thermal annealing, argon ion sputtering, ten oxidation/desorption cycles in 10^{-6} mbar oxygen atmosphere and final annealing to 2,400 K for 20 s. Surface structure and cleanliness were scrutinized via low-energy electron diffraction and Al-K α X-ray photoelectron spectroscopy. Gaseous molecular iodine was introduced into the preparation volume via sublimation of a solid piece of iodine that was thoroughly degassed using cryogenic distillation and several cycles of evaporation before the sample preparation. A saturated monolayer of iodine atoms³³ was adsorbed onto the crystal by dosing ~ 40 Langmuir of gaseous iodine. Iodine molecules dissociate upon adsorption in the first monolayer³⁴. Subsequent growth of additional layers of iodine necessitates cryogenic cooling and is unstable to NIR irradiation due to the broadband visible light absorption of molecular iodine. Saturation was verified in the range between ~ 6 and ~ 30 Langmuir by photoelectron spectroscopy with XUV light. Submonolayer (non-saturated) coverages were created via subsequent thermal desorption of iodine³⁵ at temperatures up to 1,400 K and quantified via XUV photoelectron spectroscopy. Exemplary photoelectron spectra are presented in Extended Data Fig. 1a. The background of inelastically scattered electrons in the steady-state spectra in Extended Data Fig. 1a was subtracted using a Shirley background³⁶ in energy regions of primary electron emission and a constant background elsewhere. The background subtraction scheme is illustrated in the inset of Extended Data Fig. 1a.

Electron detection. Electron momentum spectra were recorded along the XUV polarization direction via time-of-flight spectrometry. An electrostatic lens assembly was employed during streaking spectroscopy to increase the energy-dependent electron acceptance angle during time-dependent measurements. The half-angle acceptance for electrons from the He1s and I4d electron energy region was $\sim 10^\circ$ in both measurements. The half-angle acceptance for W4f electrons was $\sim 12^\circ$, and the half-angle acceptance for valence-band electrons was $\sim 22^\circ$. Both streaking and RABITT measurements can suffer from delay distortions for large angles between the electron detection and the laser field polarization direction²⁴. However, these only occur at substantially higher acceptance angles than employed in the present experiment. The energy-dependent lens transmission was corrected for before delay extraction.

Energy-averaged delay extraction. All delays reported were extracted using established algorithms¹⁰. A detailed discussion, validation and comparison to other delay extraction methods is found elsewhere¹⁵. Energy-averaged delays were retrieved by fitting the strong-field solution of the time-dependent Schrödinger equation³⁷ $P(E_f, \tau) \approx |\int_{-\infty}^{\infty} \mathcal{E}_{\text{XUV}}(t - \tau) d\phi_{\text{Volkov}}(\mathcal{E}_{\text{NIR}}, E_f, t) e^{-itE_i} e^{itE_f} dt|^2$ in the central-momentum approximation³¹ to the experimental spectrograms. The streaked photoelectron spectrum $P(E_f, \tau)$ at final kinetic energy E_f and XUV–NIR delay τ is determined by the energy of the unstreaked continuum state E_i , the dipole matrix element d , the envelope of the attosecond pulse \mathcal{E}_{XUV} and the NIR electric field \mathcal{E}_{NIR} through the Volkov phase ϕ_{Volkov} . This technique prevents loss of information as in, for example, methods treating only the first moments in energy of the spectral distributions. The robustness of the method is increased by parametrizing all light and electron pulses as Gaussians with up to second-order phase. Inelastically scattered, non-streaking (that is, XUV–NIR-delay independent) background electrons are subtracted without any assumptions about their shape by taking the derivative along the XUV–NIR delay. The static shape of the photoemission features was fixed to synchrotron data^{38,39} for stability. An individual electron wave packet group delay and chirp were fitted to each resolved spectral feature. Extended Data Fig. 2 shows a typical streaking spectrogram of the I/W(110) surface and its reconstruction.

Energy-resolved delay extraction. Energy-resolved photoemission delays were retrieved using a generalized-projections-type algorithm, similar to those employed in frequency-resolved optical gating, constructed following refs^{31,40}. The algorithm is based on the same equations and approximations as above, but employs no assumptions about the phase or amplitude of the electron wave packets. To overcome the limitations of the central momentum approximation, several retrievals are run simultaneously at the central momenta of the individual photoemission lines. The retrievals are coupled in between steps by mixing the retrieved NIR waveforms, allowing accurate photoemission delays to be retrieved. Background caused by inelastically scattered electrons was subtracted before the retrieval using the Shirley scheme detailed in Methods section 'Sample preparation and coverage calibration'.

Iodomethane/iodoethane helium sample preparation. Gas-phase experiments were carried out on iodomethane or iodoethane mixed with helium. Iodomethane and iodoethane were bought from commercial suppliers (I8507, Sigma-Aldrich; I7780, Sigma-Aldrich). The stainless-steel cylinder containing these molecules was cooled using liquid nitrogen and then evacuated to remove residual air. Several mixing cycles were performed and discarded before each measurement to

preserve purity. To maintain a stable backing pressure and helium/iodine-containing molecule ratio during the measurement, a large volume of gas mixture was prepared before the experiment. It was created by filling an evacuated reservoir up to the molecules' vapour pressure. This vapour was then diluted with helium until the desired ratio was achieved.

Overlap of molecular valence orbital electrons with He1s electrons. In our gas-phase reference measurement, the necessary spectral bandwidth of the XUV radiation causes spectral overlap of electrons emitted from the He1s and molecular orbitals. The same is true for the spectral overlap of shake-up photoemission from helium with electrons from I4d orbitals. Exemplary spectra of only iodoethane and a gas mixture are shown in Extended Data Fig. 3a. During the gas-phase experiments, the molecule/helium ratio was controlled such that both the helium shake-up states and the molecular orbitals yield only an $\sim 5\%$ contribution to the dominant I4d and He1s peaks, respectively. Retrieving the photoemission delay from simulated spectrograms containing overlapping photoemission features reveals that our retrieval method weights the photoemission delay of small spectral admixtures to a dominant feature quadratically with their intensity. Hence, the retrieval algorithm also extracts meaningful photoemission delays for the dominant part of a photoemission feature in the presence of small admixtures. Owing to their small photoemission intensity, molecular valence orbitals and helium shake-up satellites do not measurably influence the observed I4d and He1s delays, respectively. To verify this statement, we varied the ratio of helium to iodoethane in our measurements by more than a factor of four, leading to molecular orbital contributions to the helium peak of between 2% and 9%, see Extended Data Fig. 3b. We find no evidence of molecular orbital or helium shake-up influence on the measured photoemission delays, corroborating that the minor admixtures do not distort the results reported here.

Core-level photoemission from iodine on a surface and in a molecule. Not all photoemission delays from an atom in a molecule or on top of a surface are equivalent. Factors influencing delays in these different environments can be the hybridization and spectral overlap of orbitals, charge redistribution due to the chemical environment, scattering of electrons at other constituents of the molecule during their exit, alignment of the bond axis versus random bond orientation and modulation of the Coulomb–laser coupling due to screening by the metallic surface. We thus paid particular attention to the selection of our chronoscope species. This will be addressed in more detail below.

Hybridization and spectral overlap of atomic orbitals. To unambiguously extract the photoemission delay of a given atomic orbital, it should be spectrally separable from other orbitals in the spectrum. This fact prohibits the use of valence electrons as references in both stages of the experiment since their orbitals' shape and energy are not spectrally separable from surface valence electrons and, furthermore, are heavily influenced by the chemical environment. We therefore picked photoemission from the I4d core orbitals as an absolute time zero marker. Owing to its high binding energy, it is not involved in the bonding of the iodine to the surface or the molecule and is sufficiently spectrally separable from all surface and He1s photoemission features. Initial experiments were conducted using xenon adlayers, which also offer a usable 4d core level. However, all rare gases apart from He condense in islands upon physisorption, excluding the preparation of a homogeneous layer of clock atoms with variable coverage and density⁴¹.

Orbital shape and charge localization. A change in the photoemission time delay could occur if the localization of electron charge around the core or the shape of the charge density were to change substantially. The I4d orbital shapes and hybridization with carbon or tungsten orbitals can be examined using quantum chemistry packages. We used GAMESS-US^{42–44}, a triple-zeta-basis set^{45,46} and the CAM-B3LYP exchange–correlation functional⁴⁷ to calculate the I4d electron densities in different bound states. The overlap integral between corresponding I4d orbitals is $>99\%$ for iodine bound to CH₃ and W, stressing the immunity of the core orbital to external influences.

Electron scattering at the molecule and bond orientation. Further factors could influence the emission delay from the I4d orbitals: for example, the random molecular orientation with respect to the XUV radiation as compared to the aligned orientation on the surface, and the scattering of the continuum electron at the molecule during its exit. To estimate effects of both, we calculated the angle-dependent photoemission delay for I4d emission from iodomethane. For the calculations, we used GAMESS-US^{42–44}, a triple-zeta-basis set^{45,46} and the local-density approximation⁴⁸ for the initial state, ePolyScat^{49,50} for the photoionization and electron scattering calculations, and ePSProc⁵¹ for the data analysis. ePolyScat has previously been applied for the extraction of molecular photoemission delays^{23,51}. While scattering and angle dependence can have significant effects at low continuum electron energies, we find a maximum difference of less than 2 as between the orientation-averaged delay (the molecular case) and delay along the bond axis (the surface case) at 105 eV photon energy.

Giant dipole resonance properties. Molecular shape resonances arise in the absorption spectra of molecules and have been observed to contribute to the

photoemission delay²³. These are fragile and susceptible to the chemical environment, that is, they disappear when the molecular structure is changed. By contrast, the spectral properties of atomic giant dipole resonances are largely independent of the chemical environment. The *I4d* giant dipole resonance position and width in atomic iodine and in iodomethane coincide^{52,53}. Even the ionization state of atomic iodine has only minor influence, reinforcing the above statement⁵⁴. Because the phase and amplitude properties of resonances are closely linked⁵⁵, a modification of the giant dipole resonance delay properties of iodine as adsorbate or in a molecule is unlikely.

Coulomb–laser coupling and image potential screening. The streaking time determined for the chronoscope atom iodine in the gas phase contains the contribution from the Coulomb–laser coupling (CLC)¹¹. It is therefore important to inquire into the modifications due to adsorption at a metallic surface. Dynamical screening by conduction-band electrons will eventually shield the ionic core charge of the photoionized iodine, replacing the ionic potential, $-1/r$, entering CLC by an asymptotic Coulomb-like image potential, $-1/4z$. Here, r is the distance to the ion and z is the distance above the surface. This time-dependent dynamical screening potential⁵⁶ can be approximated in our exit delay simulation (see below) by a velocity (v) dependent potential experienced by the escaping electron, $V(z, v) = -\frac{1}{z} \exp\left(-\frac{\omega_s z}{v}\right) - \frac{1}{4z} \left[1 - \exp\left(-\frac{\omega_s z}{v}\right)\right]$, with ω_s the surface plasmon frequency⁵⁷. Since the dominant fraction of the CLC time shift is collected over the first few ångströms, the corrections due to dynamical screening at large distances are in the present case small. Uncertainties induced into the CLC timing by dynamical screening are found to be less than 3 as.

Gas-phase data. All individual gas-phase gauge measurements are summarized in Extended Data Fig. 3c. We find no change in the photoemission delay of *I4d* between iodoethane and iodomethane within the experimental uncertainty. This result suggests that the changes in intra-molecular photoelectron scattering and in the chemical environment do not affect the orientation-averaged photoemission delay from the *I4d* core orbitals. The increased photoemission time delay of the *I4d* orbitals compared to the He1s orbital is caused by the giant dipole resonance. It is created by a well-like effective photoelectron potential which can transiently trap the escaping electron^{23,58}.

I/W(110) data and delay dependence on coverage. All individual I/W(110) measurements are summarized in Extended Data Fig. 1b. Owing to the dielectric properties of the iodine layer, we do not expect a change in the streaking field near the surface and, hence, the *W4f* emission delay. Thus, we fit a linear function to the coverage-dependent photoemission delay results. The regression yields a *W4f* exit delay of $\Delta\tau_{W4f-I4d}(\theta) = (78 \pm 5)$ as + $\theta(8 \pm 7)$ as, which confirms near-negligible coverage dependence.

Pristine surface delay measurements. Measurements of the photoemission time delay of the W(110) surface were performed with the same XUV mirror as the iodine adsorbate and gas-phase measurements. Owing to the high reactivity of the tungsten surface, we investigated influences of surface contamination on the observed photoemission delay. Results are shown in Extended Data Fig. 1c. We find a weak increase (0.12 as min^{-1}) of the *W4f*–CB delay on the measurement time after preparation of a pristine tungsten surface. The relative photoemission time delay of the clean surface is recovered by linear extrapolation towards zero irradiation time of the W(110) crystal surface, which shows that *W4f* core-electron emission is delayed by $\tau_{W4f-I4d}^{\text{clean}} = (63 \pm 6)$ as with respect to emission from the conduction band.

DFT calculations. We employ DFT calculations using the VASP software package and the PBE exchange–correlation functional^{59–62}. We use a cut-off energy of 230 eV in line with the PAW potentials for tungsten and iodine, and a minimum of $40 \times 40 \times 1$ *k*-points. After converging these parameters and the bulk lattice constant for a bulk supercell, we consider 8 layer W(110) slabs, with up to 72 W atoms (for the 3×3 supercell). Normal to the surface, the periodic images are separated by 45 Å of vacuum. The PAW potentials include 6 (7) active electrons for W (I), resulting in up to 453 active electrons for a 3×3 slab and three iodine atoms ($72 \times 6 + 3 \times 7 = 453$). All surface geometries are optimized by allowing the iodine atoms and the top three W layers to relax.

Semiclassical transport simulations. The timing of the electron emission is simulated by combining the atomic time delay given by the EWS time delay with the time delay due to the transport from the point of photoabsorption in the crystal to the arrival at the exit surface clocked by the NIR streaking field. Initial conditions for position and momenta for the transport simulation for the classical electron ensemble are derived from DFT calculations of the electron density and the density of states for the W(110) surface. Furthermore, the spectral width of the ionizing XUV pulse is included in the initial energy (and momentum) distribution of the electron departing from the site of photoabsorption. The subsequent transport of the electron distribution (or classical wave packet) through the medium takes into account interactions with ionic cores and electrons in terms of a stochastic sequence of scattering events⁶³, the properties of which are derived from

doubly differential cross-sections for inelastic²⁶ and elastic scattering⁶⁴. Following the trajectories until the photoelectron eventually escapes from the target surface, the escape time (transition through the jellium edge located half a lattice constant above the topmost layer) is recorded. Within this simulation, enhanced photoemission from surface states can be included by adding their contribution (as modelled by DFT) to the distribution of initial conditions. Streaking spectra are modelled by switching on the vector potential at escape time and adding it to the canonical momentum. Electron trajectories outside the acceptance angle of our detector are removed from the analysis. Different starting depths, different initial energies defined by the XUV spectrum, different transport lengths within the crystal due to the random scattering distribution and varying exit angles of the photoelectrons lead to a temporal broadening of the recorded electron wave packet adding to the XUV pulse duration. The simulation allows us to directly extract the microscopic average escape time from the solid. At the same time, it allows us to determine the time shift recorded by streaking spectra. We find agreement between the two within 5 as, thereby validating the interpretation of streaking time shift as photoemission timing information.

Tungsten surface state delay. To obtain an estimate of the delay expected for electrons emitted from surface states, we performed DFT calculations of a clean W(110) surface and investigated states within the surface bandgap. Extended Data Fig. 4a shows a 2D cut of the electronic density of the surface state. Streaking starts at the jellium edge, which is located 1.1 Å above the uppermost crystal layer. Relevant dimensions are indicated in Extended Data Fig. 4a. The inelastic mean free path, that is, the average escape depth for conduction-band photoelectrons in the presented experiment is approximately 4.2 Å, which would lead to an escape delay of $\tau_{WCB} \approx 66$ as. The localization of the surface state close to the jellium edge results in an average crystal exit delay of only $\tau_{ss} \approx 11$ as. Owing to the proximity to the surface, inelastic scattering only negligibly affects photoelectrons from the surface state.

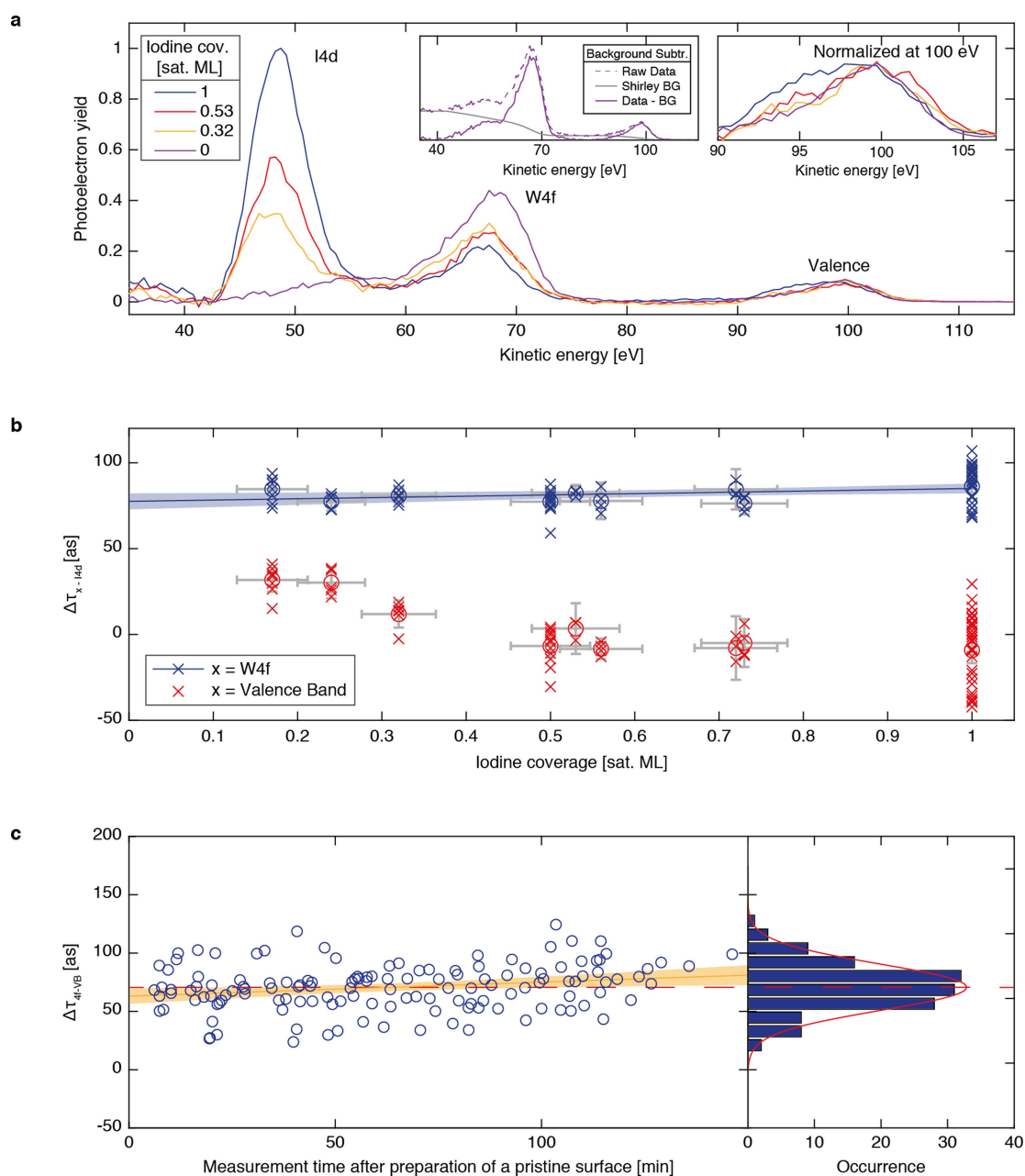
Iodine and tungsten valence states. To scrutinize the origin of valence electrons in different energy regions, we calculated the energy-resolved density of states of an iodine covered W(110) surface via DFT. Results obtained for a 2×2 W + 2I slab ($\sim 86\%$ saturated monolayer coverage) are presented in Extended Data Fig. 4b. The calculated DOS convoluted with the spectral width of the XUV source reproduces the experimental results. The density of states in the proximity of the tungsten surface atoms is mainly found in the first 3 eV below the Fermi edge E_F . Between 3 eV and 4 eV below the Fermi edge, the density of states is dominated by contributions found near the iodine adsorbates. These findings corroborate that the spectral differences observed in the experiment allow separation of the valence photoelectron feature into iodine and tungsten dominated spectral regions.

Data availability

The data that support the findings of this study are available from the corresponding authors upon request.

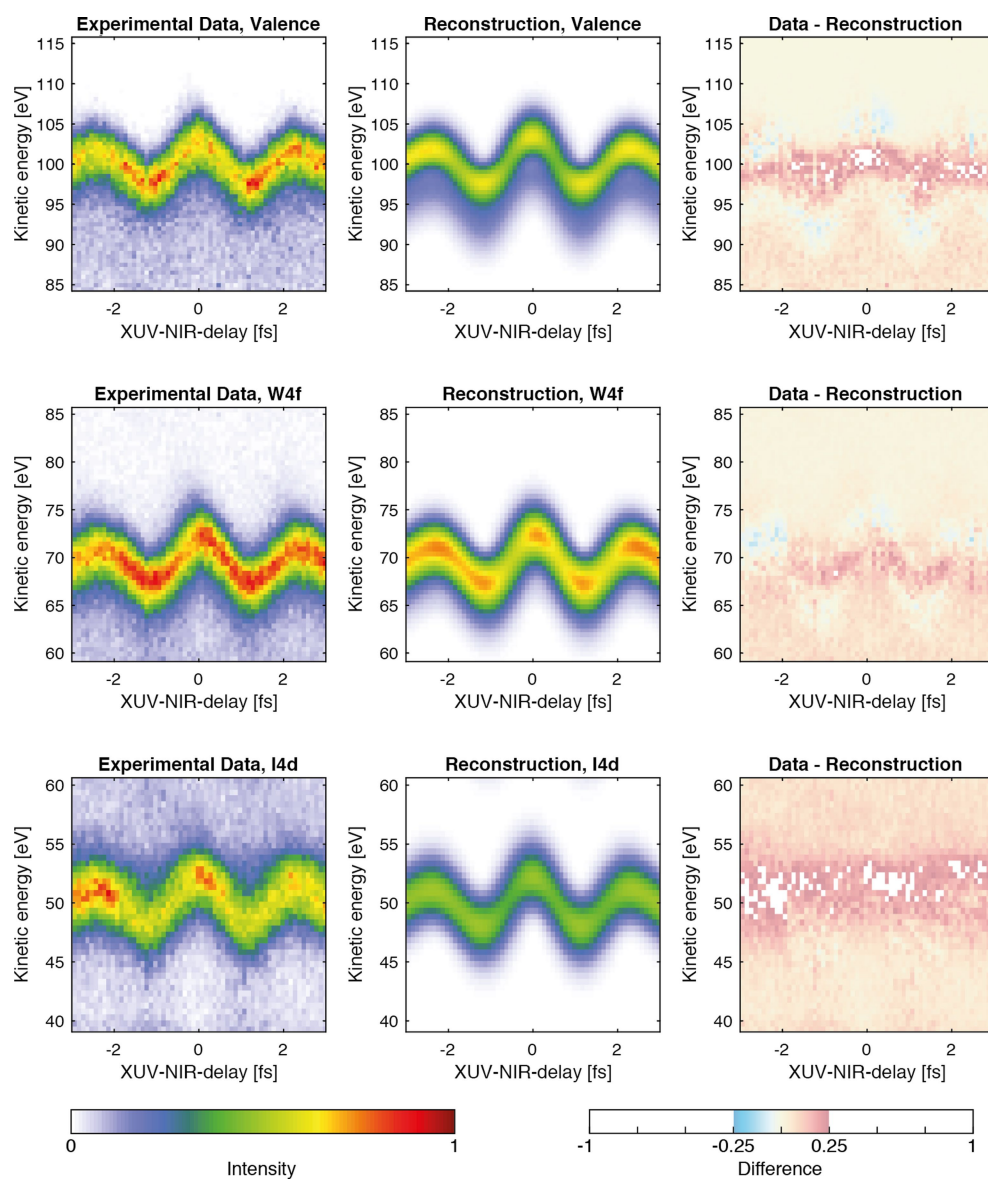
33. Jones, R. G. Halogen adsorption on solid surfaces. *Prog. Surf. Sci.* **27**, 25–160 (1988).
34. Jones, R. G. & Dowben, P. A. Reply to comments on “A re-interpretation of the LEED structures formed by iodine on W(110)” by P. A. Dowben and R. G. Jones. *Surf. Sci.* **116**, L228–L231 (1982).
35. Dowben, P. A. & Jones, R. G. A re-interpretation of the LEED structures formed by iodine on W(110). *Surf. Sci.* **105**, 334–346 (1981).
36. Shirley, D. High-resolution X-ray photoemission spectrum of the valence bands of gold. *Phys. Rev. B* **5**, 4709–4714 (1972).
37. Quéré, F., Mairesse, Y. & Itatani, J. Temporal characterization of attosecond XUV fields. *J. Mod. Opt.* **52**, 339–360 (2005).
38. Cutler, J. N., Bancroft, G. M., Sutherland, D. G. & Tan, K. H. Chemical dependence of core-level linewidths and ligand-field splittings: high-resolution core-level photoelectron spectra of *I4d* levels. *Phys. Rev. Lett.* **67**, 1531–1534 (1991).
39. Neppi, S. *Attosecond Time-Resolved Photoemission from Surfaces and Interfaces* (Technische Universität München, München, 2012).
40. Gagnon, J., Goulielmakis, E. & Yakovlev, V. S. The accurate FROG characterization of attosecond pulses from streaking measurements. *Appl. Phys. B* **92**, 25–32 (2008).
41. Dunin von Przychowski, M., Wiechert, H., Marx, G. K. L. & Schönhense, G. Real-space observation of xenon adsorption and desorption kinetics on graphite (0001) by photoemission electron microscopy. *Surf. Sci.* **541**, 46–58 (2003).
42. Schmidt, M. W. et al. General atomic and molecular electronic structure system. *J. Comput. Chem.* **14**, 1347–1363 (1993).
43. Schuchardt, K. L. et al. Basis Set Exchange: a community database for computational sciences. *J. Chem. Inf. Model.* **47**, 1045–1052 (2007).
44. Feller, D. The role of databases in support of computational chemistry calculations. *J. Comput. Chem.* **17**, 1571–1586 (1996).
45. Barbieri, P. L., Fantin, P. A. & Jorge, F. E. Gaussian basis sets of triple and quadruple zeta valence quality for correlated wave functions. *Mol. Phys.* **104**, 2945–2954 (2006).

46. Campos, C. T. & Jorge, F. E. Triple zeta quality basis sets for atoms Rb through Xe: application in CCSD(T) atomic and molecular property calculations. *Mol. Phys.* **111**, 167–173 (2013).
47. Yanai, T., Tew, D. P. & Handy, N. C. A new hybrid exchange correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **393**, 51–57 (2004).
48. Perdew, J. P. & Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B* **23**, 5048–5079 (1981).
49. Natalense, A. P. P. & Lucchese, R. R. Cross section and asymmetry parameter calculation for sulfur 1s photoionization of SF₆. *J. Chem. Phys.* **111**, 5344–5348 (1999).
50. Gianturco, F. A., Lucchese, R. R. & Sanna, N. Calculation of low energy elastic cross sections for electron-CF₄ scattering. *J. Chem. Phys.* **100**, 6464–6471 (1994).
51. Hockett, P., Frumker, E., Villeneuve, D. M. & Corkum, P. B. Time delay in molecular photoionization. *J. Phys. B* **49**, 095602 (2016).
52. Nahon, L., Svensson, A. & Morin, P. Experimental study of the 4d ionization continuum in atomic iodine by photoelectron and photoion spectroscopy. *Phys. Rev. A* **43**, 2328–2337 (1991).
53. Olney, T. N., Cooper, G. & Brion, C. Quantitative studies of the photoabsorption (4.5–488 eV) and photoionization (9–59.5 eV) of methyl iodide using dipole electron impact techniques. *Chem. Phys.* **232**, 211–237 (1998).
54. Amusia, M. Y., Cherepkov, N. A., Chernysheva, L. V. & Manson, S. T. Photoionization of atomic iodine and its ions. *Phys. Rev. A* **61**, 020701 (2000).
55. Fano, U. Effects of configuration interaction on intensities and phase shifts. *Phys. Rev.* **124**, 1866–1878 (1961).
56. Burgdörfer, J. Dynamical image charge effects on convoy electron emission from solid surfaces. *Nucl. Instrum. Methods B* **24–25**, 139–142 (1987).
57. Weaver, J. H., Olson, C. G. & Lynch, D. W. Optical properties of crystalline tungsten. *Phys. Rev. B* **12**, 1293–1297 (1975).
58. Connerade, J. P. in *Giant Resonances in Atoms, Molecules, and Solids* (eds Connerade, J. P., Esteve, J. M. & Karnatak, R. C.) 3–23 (NATO Sci. Ser. B, Vol. 151, Springer Science+Business Media, New York, 1987).
59. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
60. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
61. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
62. Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994).
63. Lemell, C., Solleder, B., Tókesi, K. & Burgdörfer, J. Simulation of attosecond streaking of electrons emitted from a tungsten surface. *Phys. Rev. A* **79**, 062901 (2009).
64. Salvat, F., Jablonski, A. & Powell, C. J. ELSEPA — Dirac partial-wave calculation of elastic scattering of electrons and positrons by atoms, positive ions and molecules. *Comput. Phys. Commun.* **165**, 157–190 (2005).



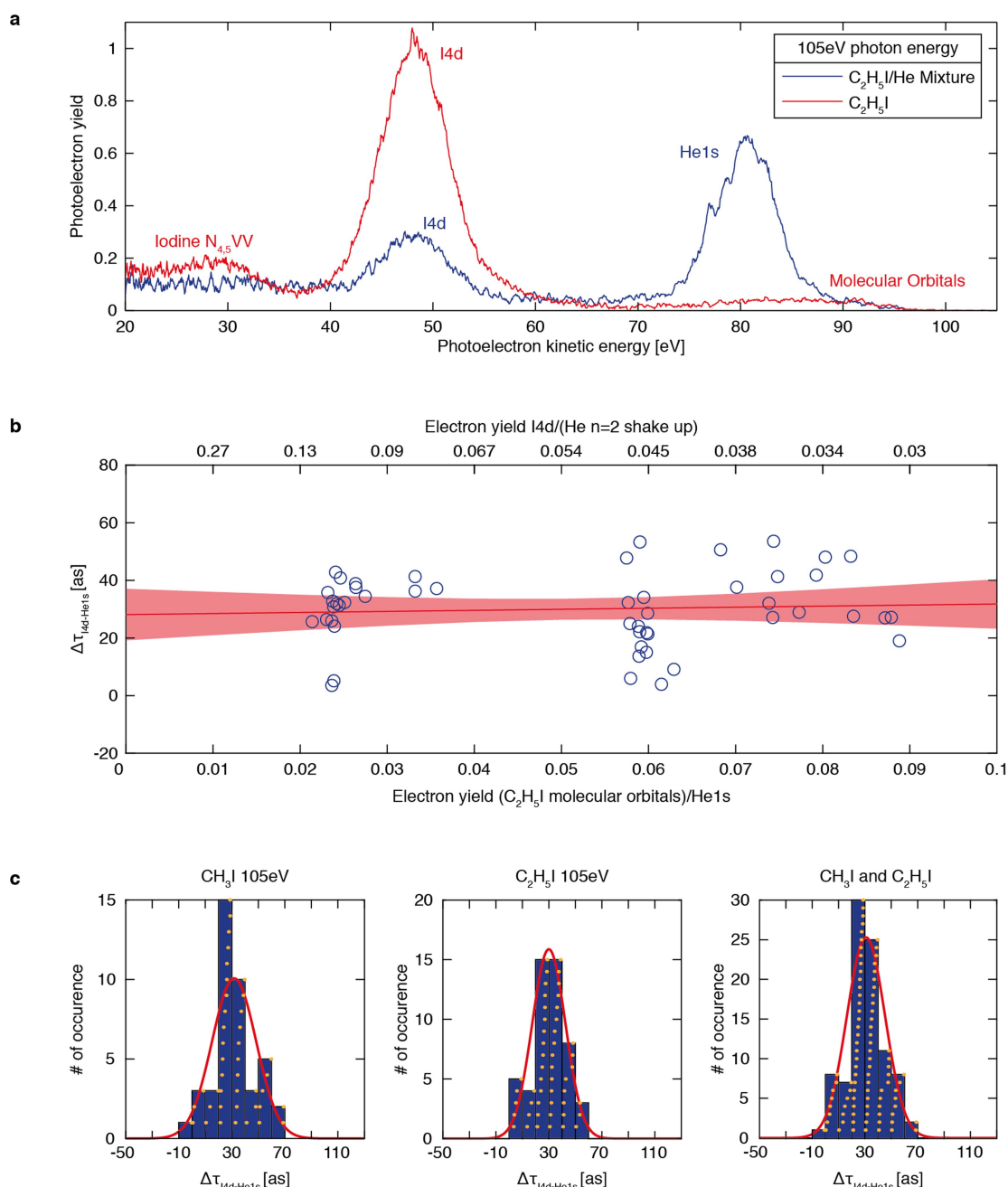
Extended Data Fig. 1 | I/W(110) and W(110) measurements. **a**, Main panel, XUV I/W(110) photoelectron spectra taken at 105 eV central photon energy for different adsorbate surface coverages (key in top left inset). Three effects are observable with decreasing iodine surface coverage: a decrease of the 14d peak intensity, an increase of the W4f peak intensity due to reduced inelastic scattering, and a shape change of the valence-electron peak towards a clean tungsten spectrum. The iodine surface density was calibrated by taking a full monolayer I/W(110) photoelectron spectrum as a reference before thermal desorption and comparing the 14d photoelectron flux before and after the thermal desorption of iodine. Because the iodine surface coverage saturates, this allows for a reliable coverage calibration. Top centre inset, illustration of the employed Shirley background (BG) subtraction scheme; top right inset, the magnified valence photoelectron spectrum. **b**, Relative photoemission delays for I/W(110). Both W4f to 14d (blue) and valence to

14d (red) delays are shown. All individual measurements are depicted by crosses, averages for individual coverages are depicted by circles. Vertical error bars mark 95% confidence assuming a Student's *t*-distribution and horizontal error bars mark maximum errors. The blue line represents a linear regression to the W4f to 14d delay (blue line), the shaded area represents the 95% confidence interval of this model. **c**, Attosecond streaking delay measurements on a pristine W(110) surface. The W4f–CB emission delay (blue circles and histogram) as a function of the time of measurement after the preparation of a pristine surface (yellow) reveals a small deviation of the centre (red, dashed line) of the normal distribution (red, solid line) fit to all measurements from the extrapolation to an instantaneous measurement due to surface contamination. The large number of measurements allows the extraction of the photoemission timing of the clean surface by extrapolation.



Extended Data Fig. 2 | Typical streaking spectrogram measured for I/W(110) at 32% of the saturated iodine surface coverage and its reconstruction. The first column shows the experimental data, the second column the results of the reconstruction algorithm and the third column their difference. The residual mainly consists of background independent of the XUV–NIR-delay, which is cancelled during the retrieval by

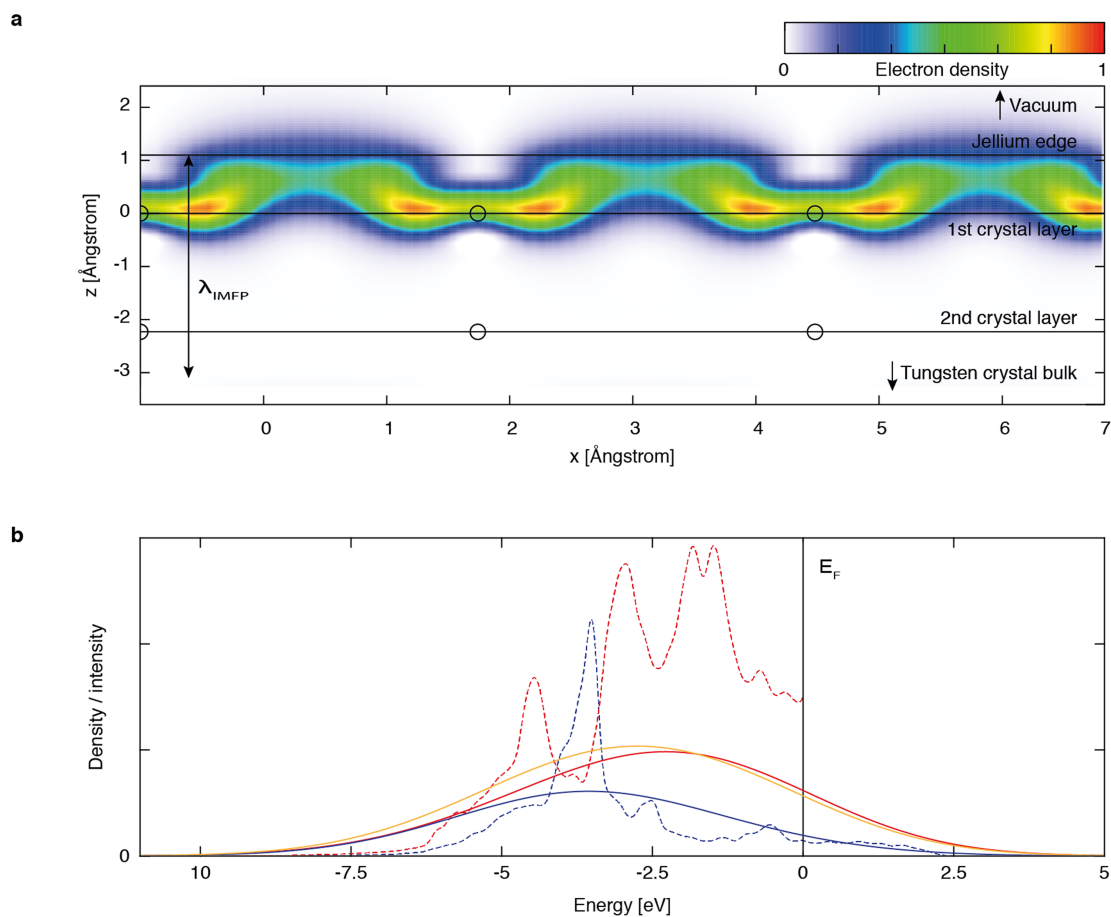
differentiating along the delay axis. The three rows focus, from top to bottom, on the streaking of the valence, W4f and I4d photoemission peaks. The XUV–NIR-delay difference between consecutive spectra in the spectrogram is 200 as. The photoemission delay results extracted from this sample spectrogram are $\Delta\tau_{W4f-I4d} = 81$ as and $\Delta\tau_{Valence-I4d} = 14$ as.



Extended Data Fig. 3 | Gas-phase iodine/helium measurements.

a, Unstreaked XUV photoelectron spectra of iodoethane (red) and an iodoethane/helium mixture (blue) recorded at 105 eV central photon energy. Electrons emitted through the $N_{4,5}VV$ Auger process are spectrally separated from all timed photoelectron peaks. **b**, Relative $I4d$ – $He1s$ photoemission delay for different mixture compositions

of iodoethane and helium. Shown are individual measurements (blue circles), linear regression (red line) and the 95% confidence interval of the regression (shaded area). **c**, Histograms (blue) of the individual relative photoemission delay measurements (yellow) between $He1s$ and $I4d$ electrons in iodomethane and iodoethane at 105 eV photon energy and normal-distribution fits to the data (red).



Extended Data Fig. 4 | Results of DFT calculations. a, DFT-derived electron density in the surface bandgap of a clean W(110) surface. A 2D cut through the (1 $\bar{1}$ 0) plane, perpendicular to the (110) surface, is shown. Tungsten atoms are indicated as black circles. Second-layer atoms are projected onto the plane of the cut. The inelastic mean free path λ_{IMFP} for conduction-band photoelectrons is marked as a guide to the eye. **b,** Energy-resolved density of states for an iodine-covered W(110) surface. The red

dashed line represents the density of states in the proximity of the top-layer tungsten atoms and the blue dashed line represents the density of states near the iodine adsorbates. The full lines are folded with the spectrum of the experimental XUV pulse. The yellow line represents the full density of states of the iodine covered tungsten surface folded with the experimental XUV spectrum. See Methods for details of ‘jellium edge’ and E_F .

Superstructures generated from truncated tetrahedral quantum dots

Yasutaka Nagaoka¹, Rui Tan¹, Ruipeng Li^{2,6}, Hua Zhu¹, Dennis Eggert^{3,4}, Yimin A. Wu⁵, Yuzi Liu⁵, Zhongwu Wang² & Ou Chen^{1*}

The assembly of uniform nanocrystal building blocks into well ordered superstructures is a fundamental strategy for the generation of meso- and macroscale metamaterials with emergent nanoscopic functionalities^{1–10}. The packing of spherical nanocrystals, which frequently adopt dense, face-centred-cubic or hexagonal-close-packed arrangements at thermodynamic equilibrium, has been much more widely studied than that of non-spherical, polyhedral nanocrystals, despite the fact that the latter have intriguing anisotropic properties resulting from the shapes of the building blocks^{11–13}. Here we report the packing of truncated tetrahedral quantum dot nanocrystals into three distinct superstructures—one-dimensional chiral tetrahelices, two-dimensional quasicrystal-approximant superlattices and three-dimensional cluster-based body-centred-cubic single supercrystals—by controlling the assembly conditions. Using techniques in real and reciprocal spaces, we successfully characterized the superstructures from their nanocrystal translational orderings down to the atomic-orientation alignments of individual quantum dots. Our packing models showed that formation of the nanocrystal superstructures is dominated by the selective facet-to-facet contact induced by the anisotropic patchiness of the tetrahedra. This study provides information about the packing of non-spherical nanocrystals into complex superstructures, and may enhance the potential of self-assembled nanocrystal metamaterials in practical applications.

Packing shapes are of interest for various disciplines, ranging from pure mathematics to industrial design, and have long been a subject of active research^{14–19}. Even the simplest Platonic shape, the tetrahedron, becomes complicated when packed in a defined space because it does not tile in the three-dimensional (3D) space of Euclidean geometry^{16,17}. After it was proposed that tetrahedra might possess the lowest packing density of any convex shape¹⁸, their arrangement into dense phases became of interest^{11,16–19}. So far, considerable progress has been made in the mathematical constructions of tetrahedral packing^{11,16–19}; this includes the seminal work on packing tetrahedra in a quasicrystalline fashion (82-tetrahedron unit cell) through exclusive, shape-induced, entropic interactions¹¹.

Unlike the extensive mathematical studies, there have been very few reports of experimental achievements in this area^{20–24}. Complex superstructures of packed tetrahedra, such as that predicted in ref. ¹¹, have not yet been observed experimentally. Here we report that three distinct superstructures—from 1D to 3D—can be obtained by self-assembly of truncated tetrahedral quantum dots (TTQDs). Although the 2D superlattices can be only tentatively assigned on the basis of the current data, our observation expands the collection of superstructures that can be constructed from tetrahedral building blocks. More importantly, our findings bring the spontaneous formation of nanocrystal assemblies to a higher level of complexity.

Monodispersed wurtzite (WZ) TTQDs were synthesized according to a previously published method with modifications²⁵. Transmission electron microscopy (TEM) revealed that the TTQDs are tetrahedral in shape with slightly truncated edges (Fig. 1a–c, Extended Data

Fig. 1). The average inorganic height of the tetrahedra is 6.7 ± 0.4 nm along the $[0002]_{\text{WZ}}$ direction (Supplementary Figs. 1, 2). The bottom $\{0002\}_{\text{WZ}}$ facet (red in the model shown in Fig. 1) is coated with octadecylphosphonic acid (ODPA), and three equivalent side $\{10\bar{1}1\}_{\text{WZ}}$ facets (blue in the model) are coated with oleic acid²⁵ (Fig. 1d, Supplementary Figs. 3–6, Supplementary Tables 1, 2). Including the surface organic ligands, the effective shape of the building blocks can be considered as a tetrahedron with an edge length l of 10.3 ± 0.4 nm (Fig. 1d).

The superstructures can be formed by drop-casting a solution of TTQDs in hexane onto a TEM grid placed on a silicon wafer (see Methods). When using relatively low particle concentrations (around 0.2 mg ml^{-1}), the predominant species formed are 1D linear TTQD assemblies (Fig. 1e) with a tetrahelical (also called a ‘Bernal spiral’) structure. The typical width of the helix ranges from 6.3 to 6.7 nm, with lengths of up to 150 nm (approximately 46 TTQDs) (Supplementary Figs. 7–12, Supplementary Tables 3–5). Figure 1f shows an example of one right-handed tetrahelix with seven interconnected TTQDs showing five consecutive atomic domains. The fast Fourier transformation (FFT) of the five atomic domains well matches the simulated FFT pattern that was based on our proposed tetrahelical model (Fig. 1f, g, Supplementary Table 3). In the model, the TTQDs are connected through defined facet-to-facet contact in a clockwise spiral (that is, $\{0002\}_{\text{WZ}}$ -to- $\{0002\}_{\text{WZ}}$ or $\{10\bar{1}1\}_{\text{WZ}}$ -to- $\{10\bar{1}1\}_{\text{WZ}}$). In addition, left-handed tetrahelices (with a counter-clockwise spiral) were also observed (Extended Data Fig. 2, Supplementary Table 5).

Controlled evaporation of a TTQD/hexane solution with a high particle concentration (around 20 mg ml^{-1}) in a glass vial led to the formation of sub-millimetre, 3D single supercrystals (Supplementary Figs. 13–15). Synchrotron-based small-angle and wide-angle X-ray scattering (SAXS and WAXS, respectively) experiments were conducted to fully elucidate the structure, with a coherence of translational alignment and orientational ordering¹³. Initially, one piece of the TTQD supercrystal was aligned in the $[001]_{\text{bcc}}$ orientation (bcc, body-centred cubic) and rotated along the $[110]_{\text{bcc}}$ axis, with a step angle of 1° (Fig. 2a). A comprehensive set of SAXS and WAXS images was collected simultaneously from the single supercrystal (Supplementary Videos 1 and 2). The SAXS data reveal a single-crystalline dotted pattern indicative of a bcc crystal phase (Supplementary Video 1). Figure 2b–m shows the SAXS and WAXS 2D images at three representative crystallographic orientations—that is, $[001]_{\text{bcc}}$, $[\bar{1}11]_{\text{bcc}}$ and $[\bar{1}10]_{\text{bcc}}$ —for rotation angles θ of 0° , 55° and 90° along the $[110]_{\text{bcc}}$ axis. The orders of rotational symmetry are the same on both the atomic scale (WAXS) and the mesoscale (SAXS); that is, four-, six- and two-fold symmetries at $[001]_{\text{bcc}}$, $[\bar{1}11]_{\text{bcc}}$ and $[\bar{1}10]_{\text{bcc}}$ projections, respectively. Notably, the calculated unit-cell lattice parameter of the obtained bcc supercrystals is 38.1 ± 0.3 nm (Extended Data Fig. 3, Supplementary Tables 6–8), which is about four times as large as the effective edge length of the TTQD building block (10.3 ± 0.4 nm). This large discrepancy suggests that the unit cell of the bcc supercrystal is

¹Department of Chemistry, Brown University, Providence, RI, USA. ²Cornell High Energy Synchrotron Source, Cornell University, Ithaca, NY, USA. ³Max Planck Institute for the Structure and Dynamics of Matter, Hamburg, Germany. ⁴Heinrich Pette Institute, Leibniz Institute for Experimental Virology, Hamburg, Germany. ⁵Center for Nanoscale Materials, Argonne National Laboratory, Argonne, IL, USA. ⁶Present address: National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, NY, USA. *e-mail: ouchen@brown.edu

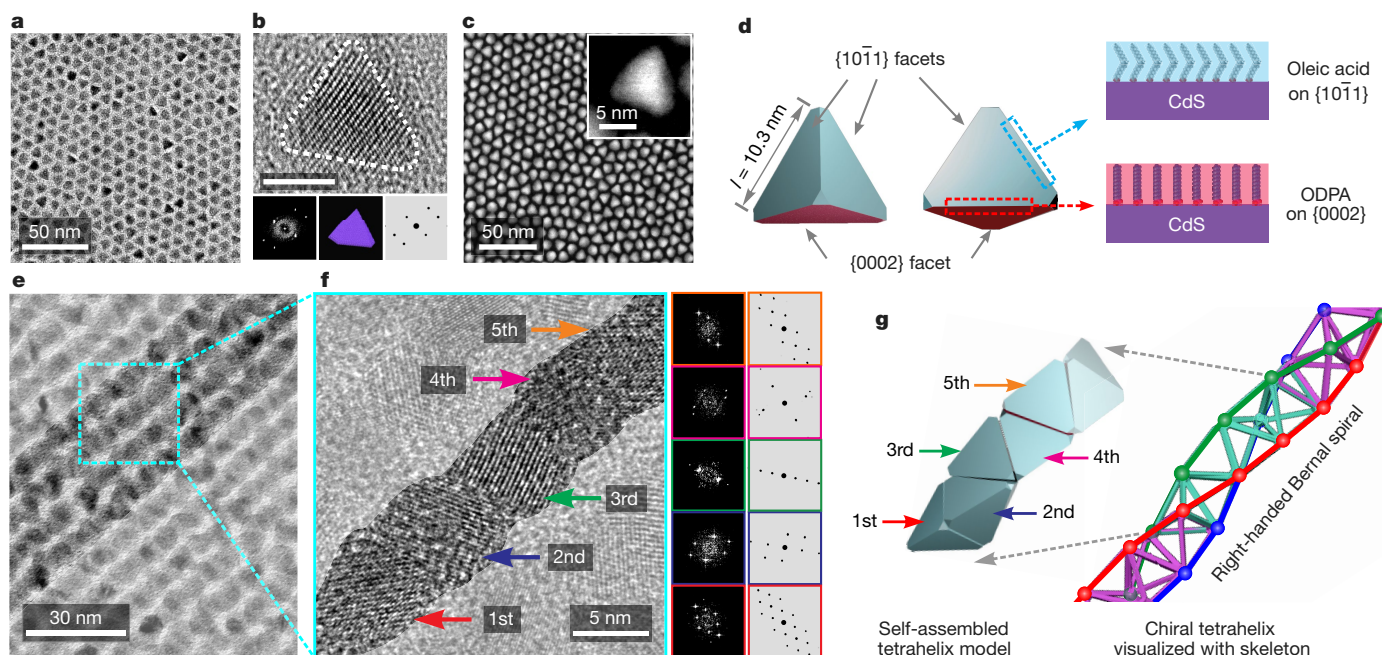


Fig. 1 | Characterization of the TTQD building blocks and tetrahedral assemblies. **a**, TEM image of TTQDs. **b**, High-resolution (HR)-TEM image (top) of an individual TTQD, the corresponding FFT pattern (bottom left) and the atomic model (bottom centre) with its corresponding simulated electron diffraction pattern (bottom right). Scale bar, 4 nm. **c**, High-angle annular dark-field imaging (HAADF)-TEM images of TTQDs. **d**, Schematic of an effective tetrahedral shape of TTQDs with an edge length of 10.3 ± 0.4 nm. The tetrahedron exhibits three $\{10\bar{1}1\}$ facets and one $\{0002\}$ facet. Oleic acid and ODPA were bound mainly on the

$\{10\bar{1}1\}$ facets (blue) and the $\{0002\}$ facet (red), respectively. **e**, TEM image of tetrahedral assemblies. **f**, HR-TEM image of a tetrahedral assembly (left), the FFT patterns corresponding to the five consecutive domains (middle) and their corresponding simulated electron diffraction patterns (right). **g**, Corresponding model of the tetrahedral assembly (left) and a skeleton visualization showing only the right-handed Bernal spiral of the chiral tetrahelix (right). We note that, owing to the overlapping of domains, only five atomic domains can be observed from seven interconnected TTQDs, as shown in **g**.

not directly formed from individual TTQDs, which was further evidenced by the computer-simulated SAXS patterns²⁶.

Recently, clathrate colloidal architectures have been assembled from DNA-modified triangular bipyramids through preformed, lower-symmetry configuration clusters⁴. In light of this finding we propose that, during the formation of the bcc supercrystals, the TTQDs first form clusters due to facet-to-facet contact ($\{0002\}_{WZ}$ -to- $\{0002\}_{WZ}$ or $\{10\bar{1}1\}_{WZ}$ -to- $\{10\bar{1}1\}_{WZ}$), and then pack into the observed bcc supercrystals at thermodynamic equilibrium^{13–15}.

On the basis of this idea, six different TTQD crystal domains were identified using a cluster model of 36 TTQDs with centering positions of $(3^4 5^{12} 6^2)$ (Fig. 2n, Extended Data Fig. 4, Supplementary Fig. 16). Each of the facet-contacted TTQD pairs with the same orientation is shown by a different colour and represents one particular crystal domain (Fig. 2n). Together, they assemble into the cluster unit with the long axis (indicated by the grey dashed-arrow in Fig. 2o) parallel with the $\{0002\}_{WZ}$ orientation of the grey TTQD pair (Fig. 2n). In addition, the rotational SAXS and WAXS measurements reveal that the long axis of the cluster is aligned in the $[110]_{bcc}$ direction, indicating a 45° offset from the $[001]_{bcc}$ Cartesian axis (Fig. 2o). Knowing the relative geometrical relationship, the unit cell of the bcc supercrystal can be constructed to be consistent with the abnormally large lattice parameter (38.1 ± 0.3 nm) calculated from the SAXS data (Fig. 2o). This proposed cluster-based bcc-supercrystal model can replicate the SAXS patterns (Supplementary Figs. 17–20) and well reproduce the WAXS signals with the corresponding rotational symmetries (along the $[110]_{bcc}$ and $[001]_{bcc}$ axes) in all projections (Fig. 2e, i, m, Extended Data Fig. 5, Supplementary Fig. 21). All other cluster models observed in the clathrate crystals, as well as simple icosahedron packing, were also examined and were ruled out on the basis of the size and crystal orientations of the cluster.

To better understand the superstructure, we fabricated a monolayer of cluster units of bcc supercrystals (Fig. 2p, Supplementary Fig. 22). The corresponding small- and wide-angle electron diffraction (SAED

and WAED) patterns are nearly identical to those of the SAXS and WAXS signals collected from the $[110]_{bcc}$ orientation (Fig. 2p, Extended Data Fig. 6), which suggests that the observed superlattice pattern is viewed along the $[110]_{bcc}$ direction. The lattice fringes exhibit a tri-line (strong–weak–weak) periodicity (Fig. 2p, Supplementary Fig. 22). The lattice distance between two adjacent strong lines, which are evenly divided by two weak lines, is 27.1 nm (Fig. 2p). This unusual tri-line-type lattice fringe is also observed at the edge of the bcc supercrystals (Extended Data Fig. 7) and can be explained by our cluster model with three-layer stacking at this projection (Supplementary Fig. 23). In addition, sequential rotations along the horizontal and vertical axes—by 34.3° and 45.0° , respectively—reveal the $[111]_{bcc}$ and $[100]_{bcc}$ orientations from the TEM images and the corresponding FFT patterns (Extended Data Fig. 8, Supplementary Fig. 24), which are consistent with the geometric relationships in a bcc crystal structure. Together, these real-space TEM observations directly verify the cluster-based bcc supercrystal model that was proposed on the basis of reciprocal space measurements from SAXS and WAXS.

We next explored other packing possibilities. When a TTQD solution with a concentration of around 2 mg ml^{-1} was evaporated, a 2D superlattice thin-film was formed with superlattice fringes of 9.4 nm (Fig. 3a, Supplementary Fig. 25). The corresponding SAED measurement reveals six-fold symmetry with a selection rule of $\{h\bar{h}00\}$ ²⁷ (Fig. 3a, Supplementary Fig. 25). Notably, the WAED exhibits a pattern with quasi-four-fold symmetry from the $(10\bar{1}0)_{WZ}$, $(10\bar{1}1)_{WZ}$ and $(11\bar{2}0)_{WZ}$ diffractions, and the $(0002)_{WZ}$ signal is absent (Fig. 3b). This symmetry discrepancy between the atomic orientation and the superstructure indicates an unusual packing of the TTQDs²⁷. Detailed analysis indicates that this superlattice can be explained by one of the previously simulated quasicrystal-approximant models with slight modifications¹¹. As in the previous model, the superlattices are assembled with two basic units: a log unit containing 12 TTQDs and a zip unit containing 20 TTQDs (Fig. 3c, Supplementary Figs. 26, 27, Supplementary Tables 9–12). We find that log- and zip units with intermediary

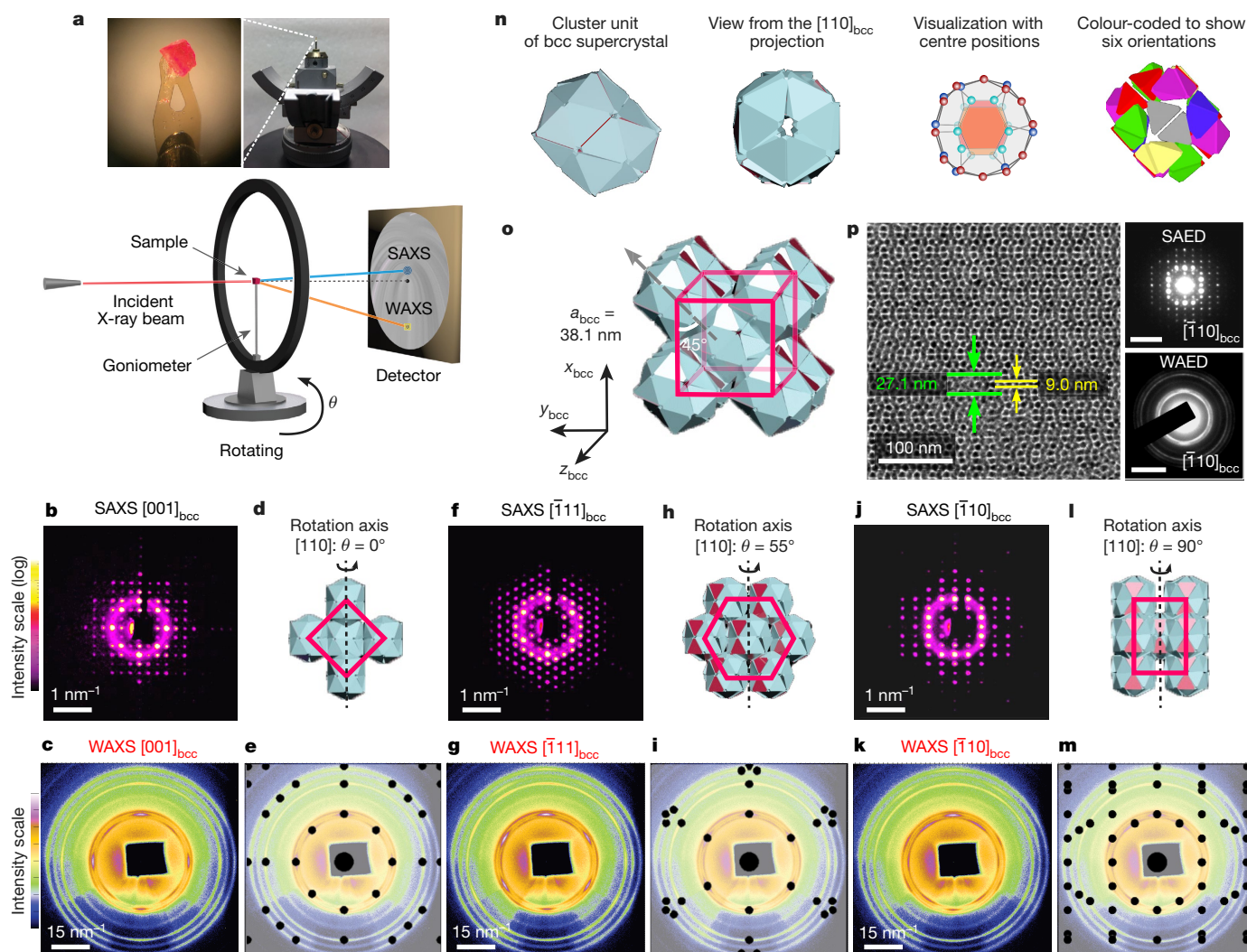


Fig. 2 | Characterization of 3D cluster-based bcc single supercrystals. **a**, Photograph of a piece of bcc supercrystal loaded on a goniometer (top) and a schematic illustration of the synchrotron-based rotational X-ray scattering setup (bottom). **b–m**, 2D images of SAXS (**b, f, j**) and WAXS (**c, g, k**) patterns at three representative crystallographic orientations: $[001]_{\text{bcc}}$ (**b–e**), $[\bar{1}\bar{1}1]_{\text{bcc}}$ (**f–i**) and $[\bar{1}10]_{\text{bcc}}$ (**j–m**). The simulated WAED patterns (**e, i, m**) were generated from computer models (**d, h, l**) at the three crystallographic orientations. **n**, Computer-generated models of a

cluster unit from the $[001]_{\text{bcc}}$ projection (left) and from the $[110]_{\text{bcc}}$ projection (middle left), a polyhedron ($3^2 4^5 6^2$) created by connecting TTQD centre points (middle right) and a cluster model with six-crystal orientation domains classified by colour (right). **o**, Computer-generated model of a unit cell of the 3D cluster-based bcc supercrystals. **p**, TEM image of a monolayer of the cluster unit viewed from the $[\bar{1}10]_{\text{bcc}}$ projection (left) and the corresponding SAED (top right; scale bar, 0.1 nm^{-1}) and WAED (bottom right; scale bar, 2 nm^{-1}) patterns.

tetrahedra can pack well via square–triangle–rhombus tiling (Fig. 3d, Supplementary Figs. 28, 29, Supplementary Tables 10–12). We note that there is no indication of the existence of pentagonal dipyramid units, as seen in the previous model¹¹. The FFT pattern of the proposed model is consistent with the FFT pattern of the TEM images and the corresponding SAED pattern (Supplementary Fig. 28). To validate our model, we conducted structural analysis of all the constituent TTQD orientations inside the log- and the zip units. WAED simulations show that the two groups of zip units inside the superlattices contribute to the simulated electron diffraction patterns with the characteristic quasi-four-fold atomic orientation symmetry and well reproduce the WAED pattern (Fig. 3e, Supplementary Fig. 29, Supplementary Tables 11, 12). No major signals from either the log unit or the intermediary tetrahedra are generated in the simulated diffraction pattern (Supplementary Fig. 29, Supplementary Table 10). In addition, we observe that the superlattice areas are exclusively assembled from either the zip- or the log units (Fig. 3f–i, Supplementary Figs. 30–34, Supplementary Tables 13–16). Figure 3f shows a TEM image with lattice fringes of 9.0 nm and a cross-fringe angle of 58.2° . This ‘zig-zag’ type of superlattice can be replicated by stacking only the zip units. Both the superlattice periodicities and the atomic electron diffraction simulation from a

zip-unit-only packing model can closely duplicate the corresponding reduced two-fold symmetries of the localized SAED and WAED patterns (Fig. 3f, g, Supplementary Fig. 30, Supplementary Table 13). Similarly, a TTQD superlattice resulting exclusively from the log units with lamellar lattice fringes (lamellar length of 9.0 nm) is also observed (Fig. 3h, i). The WAED pattern is consistent with the electron diffraction simulation without showing strongly localized $(10\bar{1}0)_{\text{WZ}}$, $(0002)_{\text{WZ}}$ and $(10\bar{1}1)_{\text{WZ}}$ signals (Fig. 3i, Supplementary Fig. 31, Supplementary Table 14). Taken together, these results demonstrate that the observed 2D superlattices that self-assemble from TTQDs may exhibit a quasicrystal-approximant packing as predicted from the thermodynamic simulation¹¹. We note that this proposed TTQD quasicrystal-approximant packing model cannot be confirmed by the current dataset owing to the high complexity of the 2D superlattices.

We note that the three superstructures assembled from TTQDs have one common structural feature: the preferred facet-to-facet alignment ($\{0002\}_{\text{WZ}}$ -to- $\{0002\}_{\text{WZ}}$ and $\{10\bar{1}1\}_{\text{WZ}}$ -to- $\{10\bar{1}1\}_{\text{WZ}}$). Additional proof of this facet-contacting preference was provided by Fourier-transform infrared spectroscopy (FTIR) (Extended Data Fig. 9, Extended Data Table 1). Compared to the amorphous TTQD powder sample, both the blue shift (about 51 cm^{-1}) of the PO_3 asymmetric stretch and the

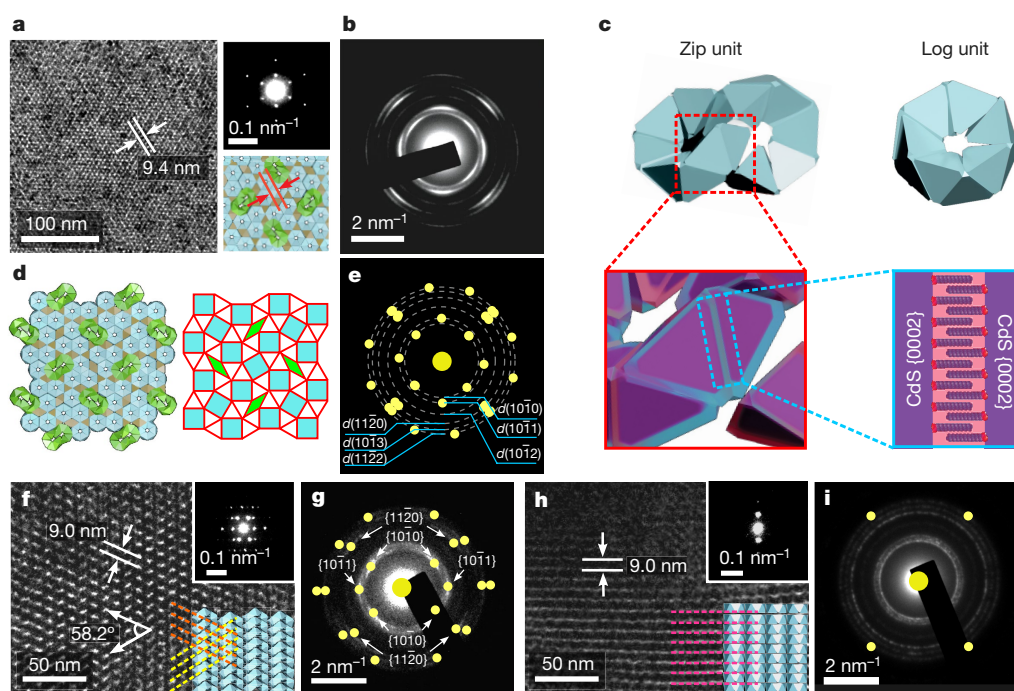


Fig. 3 | Characterization of the 2D superlattice with a tentative quasicrystal-approximant superstructure. **a**, TEM image of a 2D superlattice viewed from the top (left), the corresponding SAED pattern (top right) and a computer-generated illustration of the 2D superlattice (bottom right). **b**, Corresponding WAED pattern for the superlattice area shown in **a**. **c**, Computer-generated models of a zip unit (left) and a log unit (right). **d**, Computer-generated illustration of the 2D superlattice (left) and the tiling (right). **e**, Simulated electron diffraction pattern of the 2D superlattice model displayed in **d**. **f**, TEM image of the side view

of the assembly formed from only zip units, and the corresponding SAED pattern (inset). **g**, WAED pattern and a simulated electron diffraction pattern (yellow dots) from a zip-unit-only model. **h**, TEM image of the side view of the assembly formed from only log units, and the corresponding SAED pattern (inset). **i**, WAED pattern and a simulated electron diffraction pattern (yellow dots) from a log-unit-only model. We note that the slight deviation between the simulated electron diffraction signals and the WAED pattern probably results from small deviations in particle orientations and structural packing defects.

considerably weakened P=O vibration of the ODPA ligands from the bcc supercrystal indicate the confined state of the ODPA molecules between the two $\{0002\}_{WZ}$ facets in one TTQD pair²⁸. This confined state decreases the conformational entropy of the ligand owing to strengthened ligand–ligand interactions, thereby maximizing the structural stability at high complexities^{13,21}. It has been proven that interfacial interaction can be induced solely by directional entropic forces between hard shapes through entropic patchiness upon crowding^{6,14,15}. However, in our case, the increased anisotropic van der Waals interactions between ODPA ligands, which result in a tendency to form an ordered intermolecular structure between linear hydrocarbon chains^{29,30}, combined with the intrinsic crystal dipole along the $[0002]_{WZ}$ direction, provide additional enthalpic patchiness on the $\{0002\}_{WZ}$ facet of the TTQDs, resulting in facet alignments with specific selectivity²². As a control, when the supercrystals are formed in a more polar environment, faster supercrystal nucleation and shorter growth periods (hours rather than weeks) are observed. Consequently, instead of the cluster-based bcc supercrystal, a face-centred-cubic-like structure is generated from individual TTQDs, in which neither cluster formation nor orientation alignment of TTQDs are observed owing to the rotational freedom of the TTQDs (Supplementary Figs. 35, 36, Supplementary Table 17). This further demonstrates the role of enthalpy during the facet-to-facet contact process. In addition, facet-to-facet contact decreases the rotational freedom of the TTQDs^{13,23}, thus enabling strong atomic-orientation alignment as seen in the WAED and WAXS measurements, which ultimately enabled the otherwise impossible structural elucidation of these highly complex assemblies.

In conclusion, we found that TTQDs can self-assemble into three distinct superstructures: 1D chiral tetrahelices, 2D quasicrystal-approximant superlattices and 3D cluster-based bcc supercrystals. To our knowledge, these results provide the first experimental observation of complex superstructures assembled from single-component tetrahedral building blocks. By taking advantage of the atomic crystal phase

of the TTQDs and the unique, assembled superstructures, the details of the packing of TTQDs were elucidated using characterization techniques in both real and reciprocal spaces. In addition, we determined that entropy and enthalpy co-induced specific facet contacts, which were the major driving force for the formation of the superstructures in different dimensions. This study not only demonstrates fundamental packing strategies for tetrahedron-shaped nano-objects, but also drives momentum towards the next level of complexity in the expansion of nanocrystal assemblies across atomic, nano and macroscopic materials.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0512-5>.

Received: 18 October 2017; Accepted: 7 August 2018;
Published online 19 September 2018.

- Boles, M. A., Engel, M. & Talapin, D. V. Self-assembly of colloidal nanocrystals: from intricate structures to functional materials. *Chem. Rev.* **116**, 11220–11289 (2016).
- Feng, W. et al. Assembly of mesoscale helices with near-unity enantiomeric excess and light-matter interactions for chiral semiconductors. *Sci. Adv.* **3**, e1601159 (2017).
- Dong, A. G. et al. Binary nanocrystal superlattice membranes self-assembled at the liquid–air interface. *Nature* **466**, 474–477 (2010).
- Lin, H. X. et al. Clathrate colloidal crystals. *Science* **355**, 931–935 (2017).
- Wu, L. H. et al. High-temperature crystallization of nanocrystals into three-dimensional superlattices. *Nature* **548**, 197–201 (2017).
- Boneschanscher, M. P. et al. Long-range orientation and atomic attachment of nanocrystals in 2D honeycomb superlattices. *Science* **344**, 1377–1380 (2014).
- Liu, W. Y. et al. Diamond family of nanoparticle superlattices. *Science* **351**, 582–586 (2016).
- Weidman, M. C., Smilgies, D. M. & Tisdale, W. A. Kinetics of the self-assembly of nanocrystal superlattices measured by real-time in situ X-ray scattering. *Nat. Mater.* **15**, 775–781 (2016).
- Wang, T. et al. Self-assembled colloidal superparticles from nanorods. *Science* **338**, 358–363 (2012).

10. Cabane, B. et al. Hiding in plain view: Colloidal self-assembly from polydisperse populations. *Phys. Rev. Lett.* **116**, 208001 (2016).
11. Haji-Akbari, A. et al. Disordered, quasicrystalline and crystalline phases of densely packed tetrahedra. *Nature* **462**, 773–777 (2009).
12. Gong, J. X. et al. Shape-dependent ordering of gold nanocrystals into large-scale superlattices. *Nat. Commun.* **8**, 14038 (2017).
13. Li, R. et al. Competing interactions between various entropic forces toward assembly of Pt₃Ni octahedra into a body-centered cubic superlattice. *Nano Lett.* **16**, 2792–2799 (2016).
14. Manoharan, V. N. Colloidal matter: Packing, geometry, and entropy. *Science* **349**, 1253751 (2015).
15. Petukhov, A., Tuinier, R. & Vroege, G. Entropic patchiness: Effects of colloid shape and depletion. *Curr. Opin. Colloid Interface Sci.* **30**, 54–61 (2017).
16. Kallus, Y. & Elser, V. Dense-packing crystal structures of physical tetrahedra. *Phys. Rev. E* **83**, 036703 (2011).
17. Torquato, S. & Jiao, Y. Dense packings of the Platonic and Archimedean solids. *Nature* **460**, 876–879 (2009).
18. Conway, J. H. & Torquato, S. Packing, tiling, and covering with tetrahedra. *Proc. Natl Acad. Sci. USA* **103**, 10612–10617 (2006).
19. Chen, E. R., Engel, M. & Glotzer, S. C. Dense crystalline dimer packings of regular tetrahedra. *Discrete Comput. Geom.* **44**, 253–280 (2010).
20. Yang, M. et al. Self-assembly of nanoparticles into biomimetic capsid-like nanoshells. *Nat. Chem.* **9**, 287–294 (2017).
21. Huang, M. J. et al. Selective assemblies of giant tetrahedra via precisely controlled positional interactions. *Science* **348**, 424–428 (2015).
22. Tang, Z., Kotov, N. A. & Giersig, M. Spontaneous organization of single CdTe nanoparticles into luminescent nanowires. *Science* **297**, 237–240 (2002).
23. Boles, M. A. & Talapin, D. V. Self-assembly of tetrahedral CdSe nanocrystals: effective “patchiness” via anisotropic steric interaction. *J. Am. Chem. Soc.* **136**, 5868–5871 (2014).
24. Ghosh, S. et al. Pyramid-shaped wurtzite CdSe nanocrystals with inverted polarity. *ACS Nano* **9**, 8537–8546 (2015).
25. Tan, R. et al. Monodisperse hexagonal pyramidal and bipyramidal wurtzite CdSe–CdS core–shell nanocrystals. *Chem. Mater.* **29**, 4097–4108 (2017).
26. Förster, S. et al. Order causes secondary Bragg peaks in soft materials. *Nat. Mater.* **6**, 888–893 (2007).
27. Hahn, T. *International tables for crystallography* 5th edn, Vol. A, 1–905 (Springer, Dordrecht, 2005).
28. Wan, Y. et al. Enhanced tribology durability of a self-assembled monolayer of alkylphosphonic acid on a textured copper substrate. *Appl. Surf. Sci.* **259**, 147–152 (2012).
29. Grzelczak, M., Vermant, J., Furst, E. M. & Liz-Marzan, L. M. Directed self-assembly of nanoparticles. *ACS Nano* **4**, 3591–3605 (2010).
30. Vaia, R. A., Teukolsky, R. K. & Giannelis, E. P. Interlayer structure and molecular environment of alkylammonium layered silicates. *Chem. Mater.* **6**, 1017–1022 (1994).

Acknowledgements O.C. acknowledges support from the Brown University Startup Fund, the Salomon Award Fund, the IMNI Seed Fund and the UAC grant from the Xerox foundation. The Cornell High Energy Synchrotron Source was supported by the NSF award DMR-1332208. This work was performed, in part, at the Center for Nanoscale Materials, a US Department of Energy Office of Science User Facility, and supported by the US Department of Energy, Office of Science, under contract number DE-AC02-06CH11357. The TEM and SEM measurements were performed at the Electron Microscopy Facility in the Institute for Molecular and Nanoscale Innovation (IMNI) at Brown University.

Reviewer information *Nature* thanks J. Fang, A. Petukhov and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.N., R.L., R.T., Z.W. and O.C. conceived and designed the experiments. R.T. performed nanocrystal synthesis. Y.N. and R.T. conducted TTQD superstructure formations. R.L. and Z.W. carried out the rotational SAXS and WAXS measurements. Y.N., D.E., Y.A.W. and Y.L. performed the electron microscopy measurements. Y.N., H.Z. and R.L. conducted the data analysis and simulation. O.C. supervised the entire project. Y.N. and O.C. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0512-5>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0512-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to O.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Synthesis of CdSe core quantum dots. The synthesis of CdSe core quantum dots followed a previously reported hot-injection method³¹. 120 mg CdO (99.998%), 560 mg ODPa (99%) and 6 g trioctylphosphine oxide (TOPO, 99%) were loaded into a 100 ml flask. The mixture was degassed and heated to 150 °C for 1 h under vacuum. The reaction solution was then heated to 360 °C under nitrogen to form a clear, colourless solution. After adding 4.0 ml trioctylphosphine (TOP, 97%), the temperature was brought up to 380 °C and a freshly prepared Se/TOP (120 mg Se in 1.0 ml TOP) solution was swiftly injected into the flask. When the CdSe core reached the desired size, the reaction was quenched by removing the heating mantle and blowing the outside of the flask with cool air. The CdSe core size was estimated to be 2.7 nm from the first absorption peak. The resulting CdSe core quantum dots were stored in hexane for use in the next step.

Synthesis of CdSe–CdS core–shell quantum dots with a truncated tetrahedral shape. The quantum dots with edge-truncated tetrahedral shapes were synthesized using a CdS-shell growth protocol we developed recently with minor modifications^{25,32}. In a typical reaction, 100 nmol of CdSe cores were purified once by washing with acetone. The nanocrystals were loaded into a 100 ml three-neck flask with a solvent mixture of 2 ml 1-octadecene (ODE, 90%) and 2 ml oleylamine (OAm, 70%). The reaction mixture was degassed under vacuum at room temperature for 1 h and 120 °C for 10 min to remove hexane, water and oxygen. The reaction mixture was then heated to 310 °C under nitrogen for shell growth. When the temperature reached 240 °C, solutions of Cd-oleate (4-monolayer equivalent of CdS shell) and 1-octanethiol (1.2 equivalent of Cd-oleate), each dissolved in 2 ml of ODE were simultaneously added dropwise using a syringe pump at an injection rate of 2 ml h^{−1}. One hour after the injection was complete, the reaction was stopped by removing the heating mantle and cooling to room temperature by blowing the outside of the flask with cool air. The product was purified by three rounds of precipitation and redispersion using acetone/methanol and hexane. The particles were finally suspended in around 2 ml of hexane for the fabrication of superlattices.

Formation of nanocrystal superstructures from TTQDs. The nanocrystal superstructures were created through solvent (hexane) evaporation. Various nanocrystal superstructures were made by tuning the evaporation speed and the concentration of the nanocrystals.

1D tetrahedral assemblies were created from a 0.2 mg ml^{−1} solution of TTQD in hexane. One drop (around 20 µl) of the solution was dropped on a TEM grid placed on a piece of silicon wafer. The evaporation was complete within few seconds.

2D superlattice assemblies were created from a 2.0 mg ml^{−1} solution of TTQD in toluene. One drop (around 20 µl) of the solution was dropped on a TEM grid placed on a piece of silicon wafer and the evaporation was complete in around 30 s.

3D bcc-supercrystal assemblies were created from a 20 mg ml^{−1} solution of TTQD in hexane through slow evaporation. We placed 5 ml of solution in a 20-ml capped glass vial. The evaporation was complete in two weeks.

Monolayer superlattices of a bcc supercrystal were created from a 20 mg ml^{−1} solution of TTQD in hexane through slow evaporation. We placed a TEM grid on the bottom of a 5-ml glass vial, 0.1 ml of solution was added and the vial was capped. The evaporation was complete in around 24 h.

Interface assemblies of TTQDs were created through a slow destabilization method³³. 2 ml of a TTQD solution in hexane was placed in a 5 ml glass vial and 2 ml of ethanol was slowly added. The layers of the polar phase (ethanol) and the nonpolar phase (TTQD/hexane solution) were initially clearly separated, but slowly merged together to form one liquid phase through interfacial diffusion (several hours). A silicon chip can also be placed inside the vials as a substrate for growing the supercrystal.

Characterization techniques. Ultraviolet–visible absorption spectra were measured using an Agilent Technologies Cary 5000 UV–Vis–NIR Spectrophotometer. Typically, the samples were dissolved in hexane for the measurements.

Photoluminescence measurements were performed using an Edinburgh Instruments Fluorescence Spectrometer FS5. Typically, the samples were dissolved in hexane for these measurements.

TEM measurements were performed on a JEOL-2100F operated at 200 kV and a FEI-Philips CM20 operated at 200 kV. For the solution samples, such as the TTQD building blocks, one drop of the hexane solution with fine concentration adjustment was dropped onto a 300-mesh copper TEM grid placed on filter paper and dried under ambient conditions. 1D tetrahedral, 2D superlattice and 3D bcc-supercrystal assemblies were formed on TEM grids placed on a silicon wafer or in a glass vial. For the 3D bcc-supercrystal sample, we carefully selected a small

piece of the sample, gently placed it on a TEM grid and sliced it into thinner pieces. We note that the bcc-supercrystal samples were fragile and extra care was needed.

WAED and SAED measurements were carried out on a JEOL-2100F operated at 200 kV. Camera lengths of 20 cm and 200 cm were typically used for WAED and SAED, respectively.

HAADF scanning TEM (STEM) and STEM–energy dispersive X-ray spectroscopy (EDS) mapping was performed on a FEI Talos F200X TEM/STEM running at 200 kV equipped with a SuperX EDS detector. SEM measurements were performed on a LEO 1530 operated at 3 kV.

FTIR measurements were performed on a Jasco FT/IR 4100. The solution sample was directly dropped, or the solid sample was placed onto a NaCl pellet.

Synchrotron-based SAXS and WAXS measurements were performed at the B1 station of the Cornell High Energy Synchrotron Source (CHESS), Cornell University. Using a double circular pinhole aligned tube, monochromatic X-rays at a collimated energy of 25.514 keV were reduced to a small beam with a diameter of 100 microns. SAXS and WAXS images were collected simultaneously using a large area Mar345 detector. A mixture of CeO₂ powders was used to calibrate the sample-to-detector distances and associated detector seating parameters.

A single bcc supercrystal was loaded onto a MiTeGen mesh grid, which was subsequently mounted on the home-made two-circle rotating diffractometer. The bcc supercrystal was aligned parallel to a desirable rotation axis. Upon X-ray illumination, a series of SAXS and WAXS images were collected during sample rotation. An angular rotation step of 1° was used, and the full dataset was collected by rotation of θ over an angular range of 180°.

Analysis software. For analysis, a Fit2D program (www.esrf.eu/computing/scientific/FIT2D) was used to integrate the collected SAXS and WAXS images into 1D patterns. A Multi-Peak Fitting 2.0 package in Igor Pro version 6.37 (WaveMetrics) was used for the XRD peak analyses. CrystalMaker version 9.2 (CrystalMaker Software Ltd) was used for simulation of X-ray and electron diffraction results. Computer model construction was carried out using Autodesk 3dsMax (Autodesk). ImageJ (<https://imagej.nih.gov/ij/>) was used for 2D-FFT analysis. The WAED and WAXS patterns of the 2D superlattice and 3D bcc supercrystal were simulated using the ‘Single-electron diffraction’ function in CrystalMaker. Typically, the observed WAED and WAXS patterns resulted from multiple crystal domains. Diffraction patterns for each domain were simulated using the ‘Single-electron diffraction’ function in CrystalMaker and the resulting patterns were combined using Adobe Photoshop based on determined crystal orientations and viewing directions.

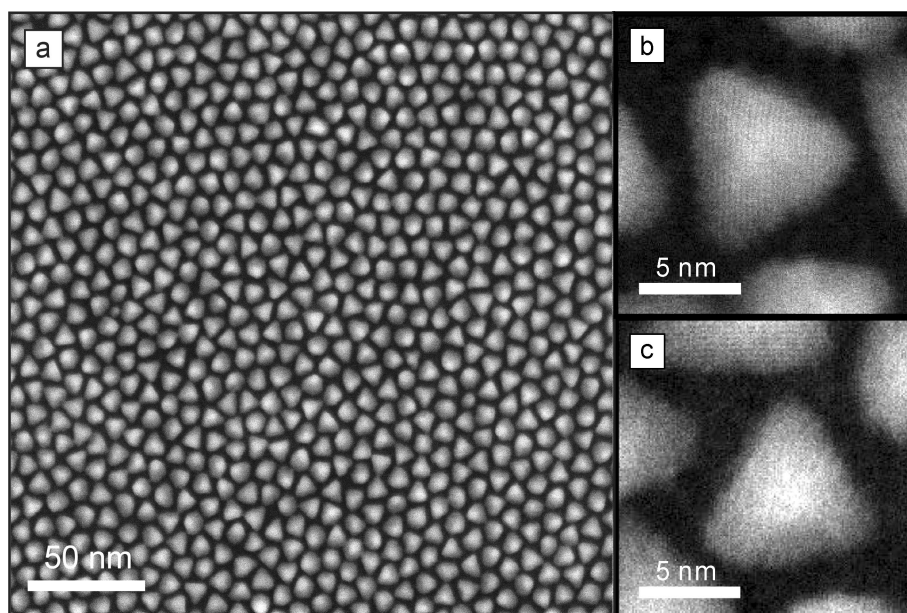
SAXS patterns of the bcc supercrystal were simulated using two methods (see Supplementary Information for details): method (i), 2D-FFT from computer-generated models of the superstructure; and method (ii), 3D-mesh pixel-based calculation of the multiplication of the form factor and the structure (lattice) factor. In method (i), a computer model of the bcc-supercrystal structure was first constructed using Autodesk 3dsMax. Smaller building blocks were used for obtaining 2D-FFT images. The computer model of the bcc-supercrystal structure was rendered from multiple projections corresponding to the observed SAXS patterns such as the [100]_{bcc}, the $\bar{1}11$ _{bcc} and the $\bar{1}10$ _{bcc} projections. 2D-FFT using ImageJ was applied to these superlattice model images. The obtained FFT patterns were further processed for ease of visualization using the ‘Gaussian Blur’ function in Adobe Photoshop, in which the FFT spots were broadened while maintaining the intensity profiles of each signal. For method (ii), we used a 2D meshgrid method and calculated an intensity for each pixel for the simulation of the 2D SAXS pattern. The details were described in the Supplementary Information section entitled ‘SAXS simulation Method (ii): Multiplication of the form factor and the structure (lattice) factor’.

Sample size. No statistical methods were used to predetermine sample size.

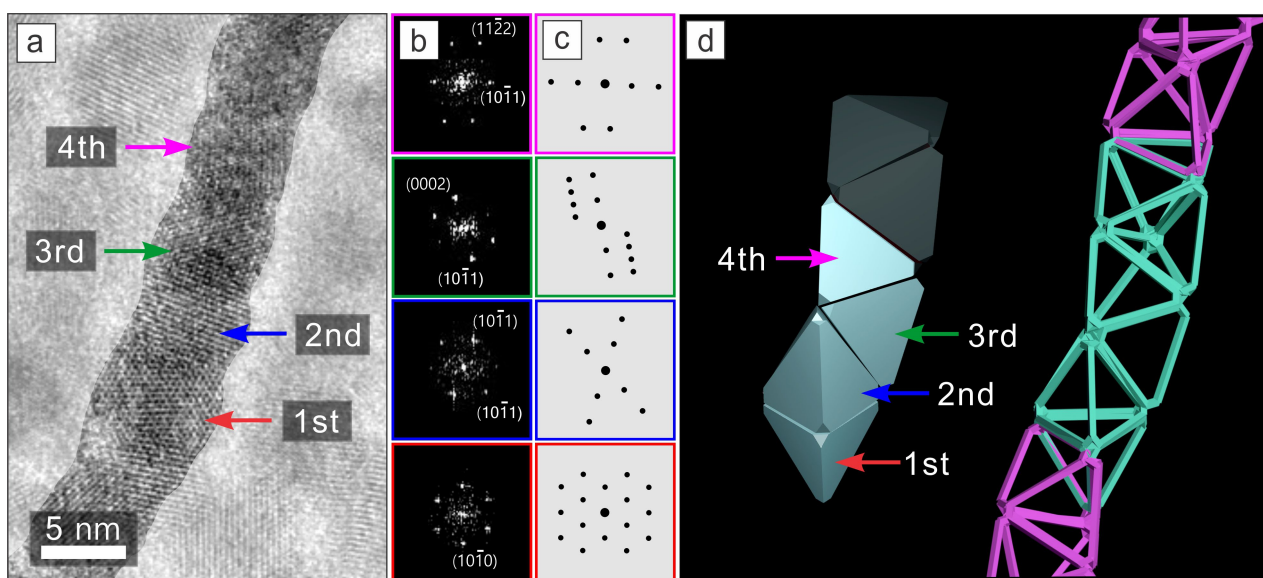
Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

- Carbone, L. et al. Synthesis and micrometer-scale assembly of colloidal CdSe/CdS nanorods prepared by a seeded growth approach. *Nano Lett.* **7**, 2942–2950 (2007).
- Chen, O. et al. Compact high-quality CdSe–CdS core–shell nanocrystals with narrow emission linewidths and suppressed blinking. *Nat. Mater.* **12**, 445–451 (2013).
- Talapin, D. V. et al. CdSe and CdSe/CdS nanorod solids. *J. Am. Chem. Soc.* **126**, 12984–12988 (2004).

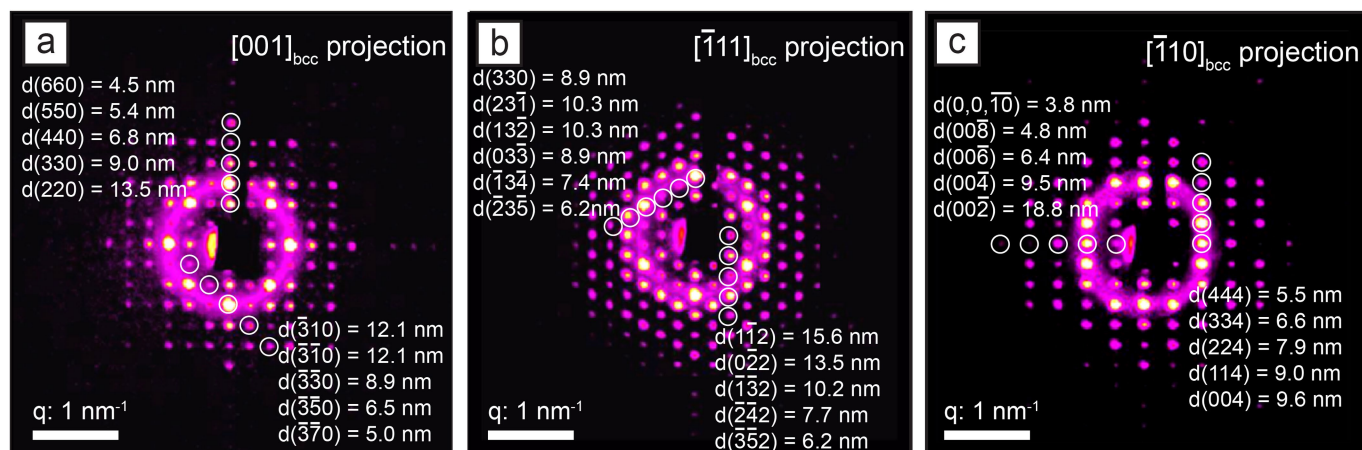


Extended Data Fig. 1 | HAADF-STEM images. a–c, HAADF-STEM images of TTQDs at low (a) and high (b, c) magnification.



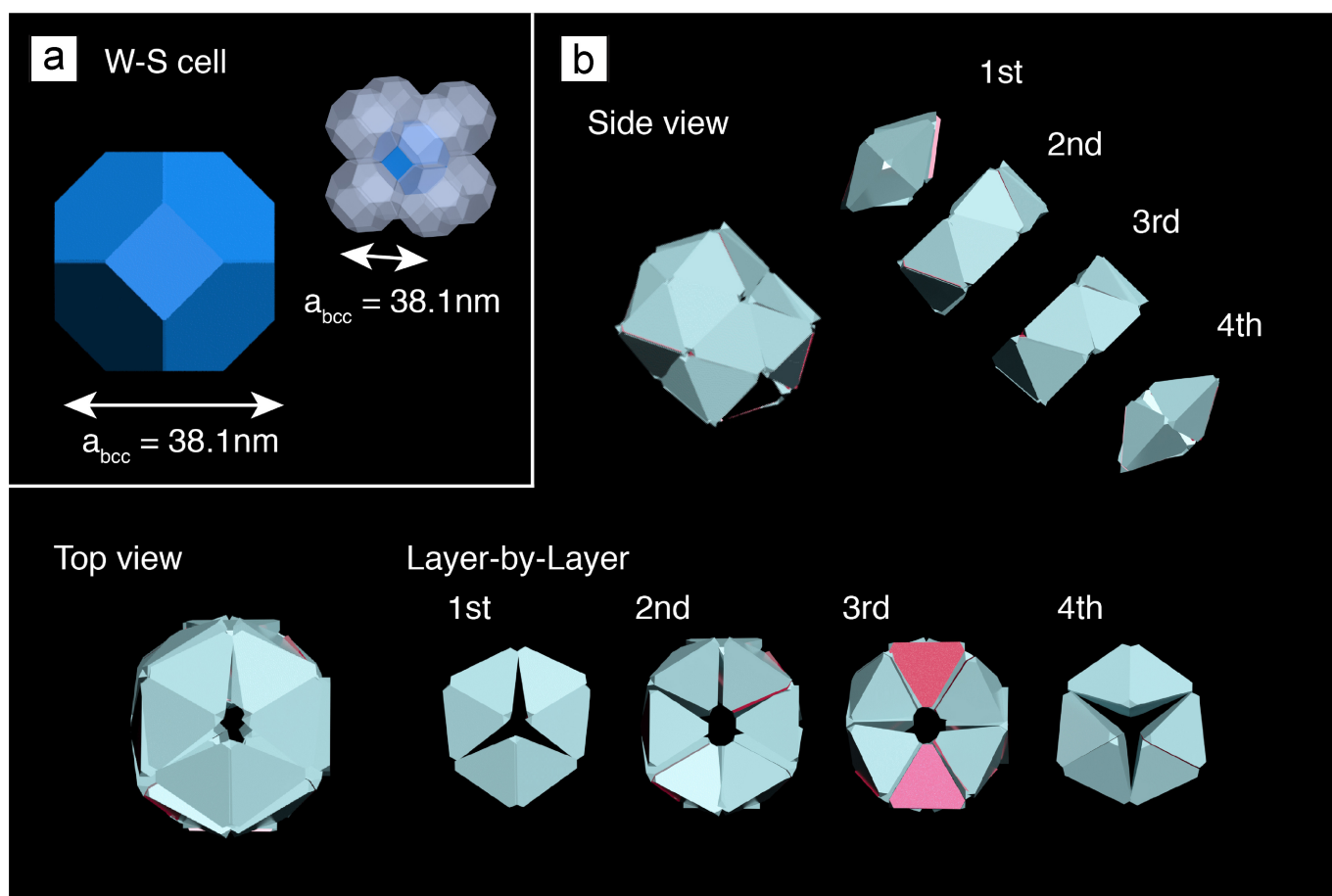
Extended Data Fig. 2 | Tetrahelix with a left-handed (counter-clockwise) spiral. **a**, HR-TEM image of a tetrahelical assembly. **b**, FFT patterns corresponding to the four consecutive domains.

c, The corresponding simulated electron diffraction patterns of the four domains. **d**, The corresponding schematic illustration of the tetrahelix.



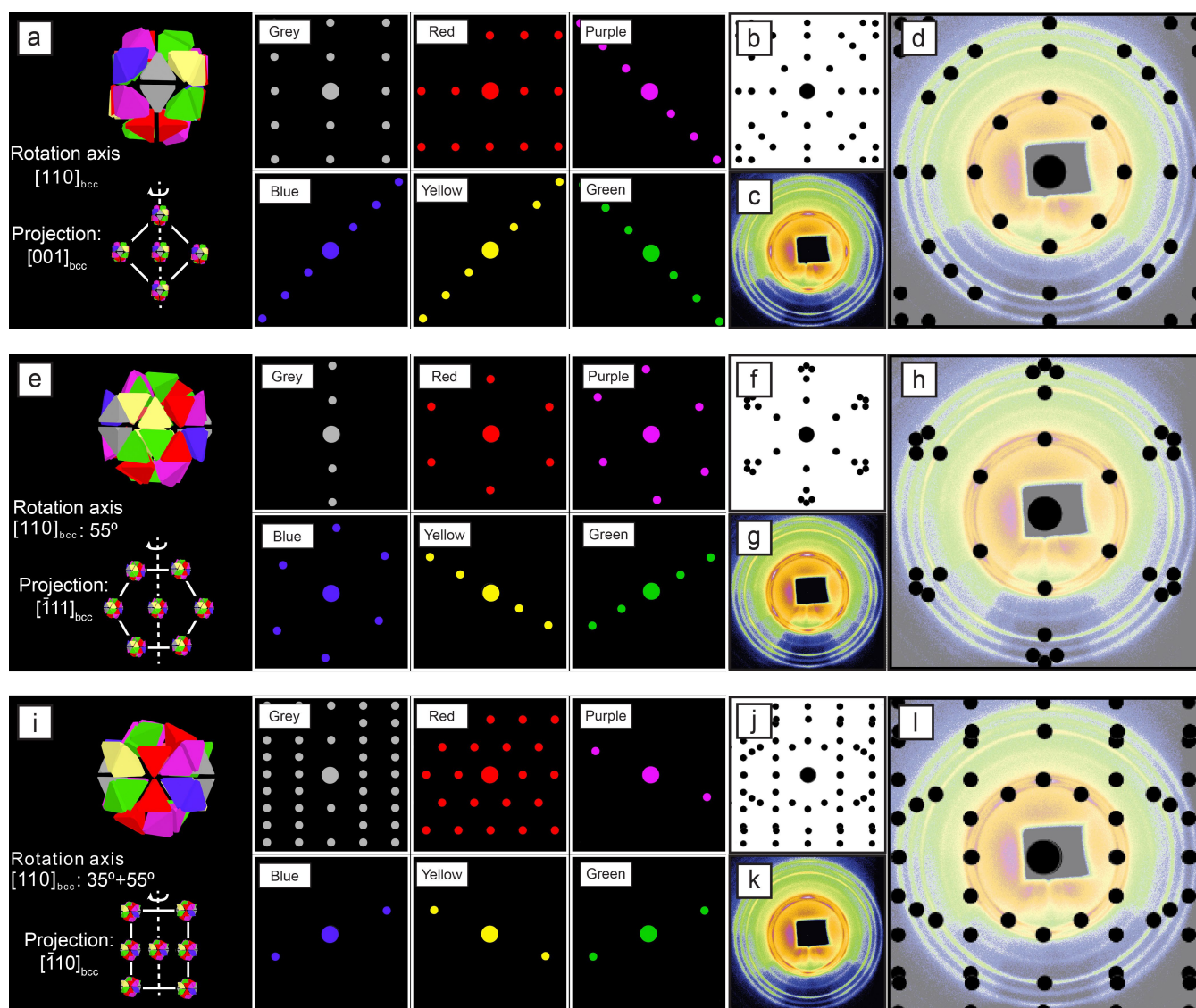
Extended Data Fig. 3 | Representative SAXS patterns from different projections obtained from rotational SAXS measurements along the

$[\bar{1}10]_{\text{bcc}}$ axis. a–c, $[001]_{\text{bcc}}$ projection (a), $[\bar{1}11]_{\text{bcc}}$ projection (b) and $[\bar{1}10]_{\text{bcc}}$ projection (c). The d -spacings of representative spots are labelled.



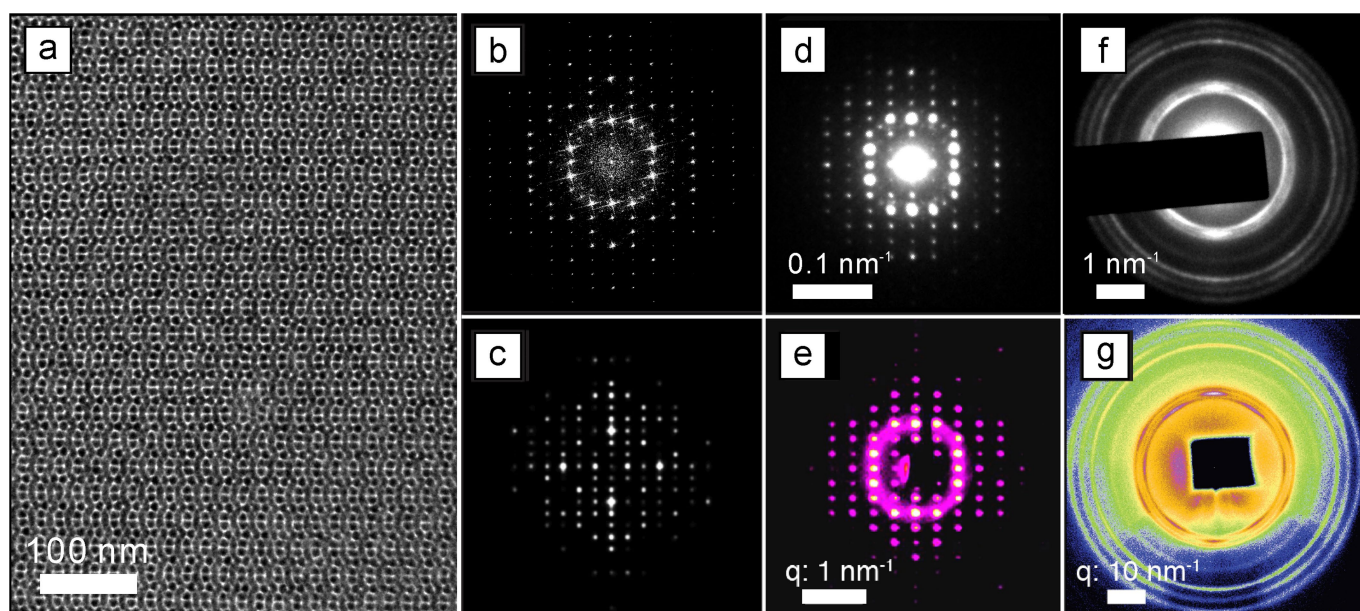
Extended Data Fig. 4 | Construction of a cluster-unit building block of the bcc-supercrystal solid from 36 TTQDs. a, Wigner-Seitz (W-S) cell of a bcc crystal structure. **b,** Schematic illustrations of the assembled and disassembled ‘ball-like’ cluster-units comprising 36 TTQDs, viewed

from the side and the top. We note that ‘bcc’ represents another level of ‘cluster-based bcc superlattice’, and if a less dense and open structure is considered, a slight symmetry-breaking of the cluster does not modify the bcc structure.



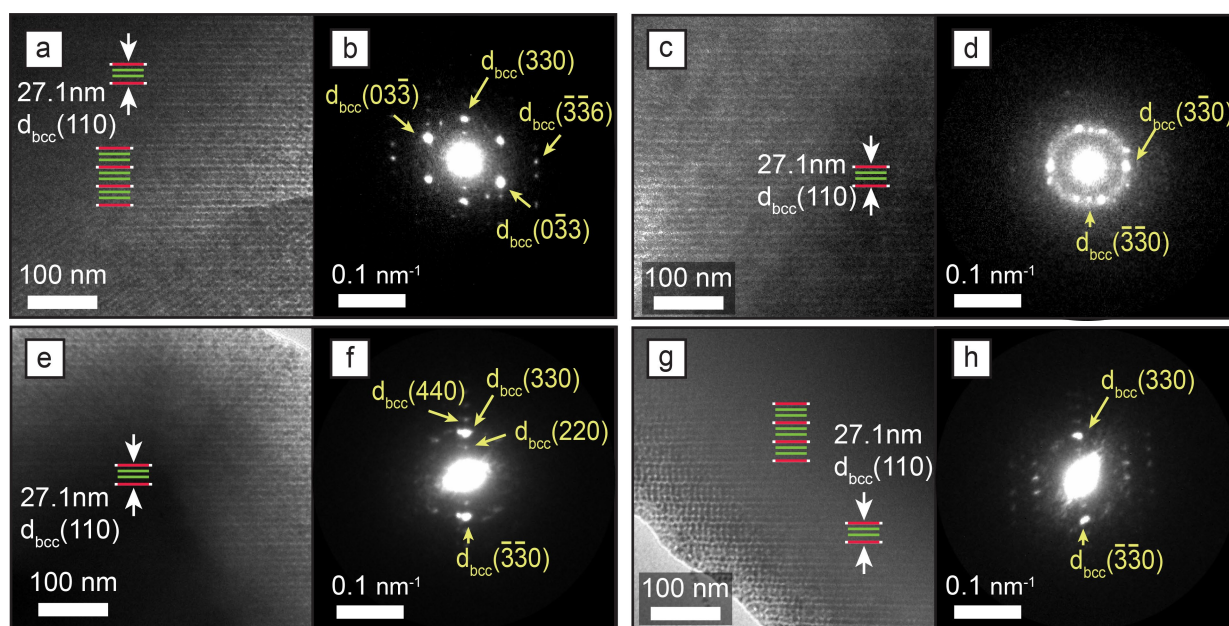
Extended Data Fig. 5 | Simulations of the WAXS patterns obtained from the rotational WAXS measurements along the $[110]_{\text{bcc}}$ axis. **a–l**, Simulations of the WAXS patterns for the $[001]_{\text{bcc}}$ projection (**a–d**), the $[\bar{1}11]_{\text{bcc}}$ projection (**e–h**) and the $[\bar{1}10]_{\text{bcc}}$ projection (**i–l**). In the cluster unit, six atomic orientations were used which are colour-coded in the computer-generated models. The orientations are classified using grey,

red, purple, blue, yellow and green colours in the model (leftmost panels in **a**, **e**, **i**) and in the corresponding simulated patterns for each orientation (right panels in **a**, **e**, **i**). Simulated WAXS patterns were generated by overlapping six orientations (**b**, **f**, **j**). The experimentally observed WAXS patterns (**c**, **g**, **k**) were compared with the resulting simulated patterns (**d**, **h**, **l**).



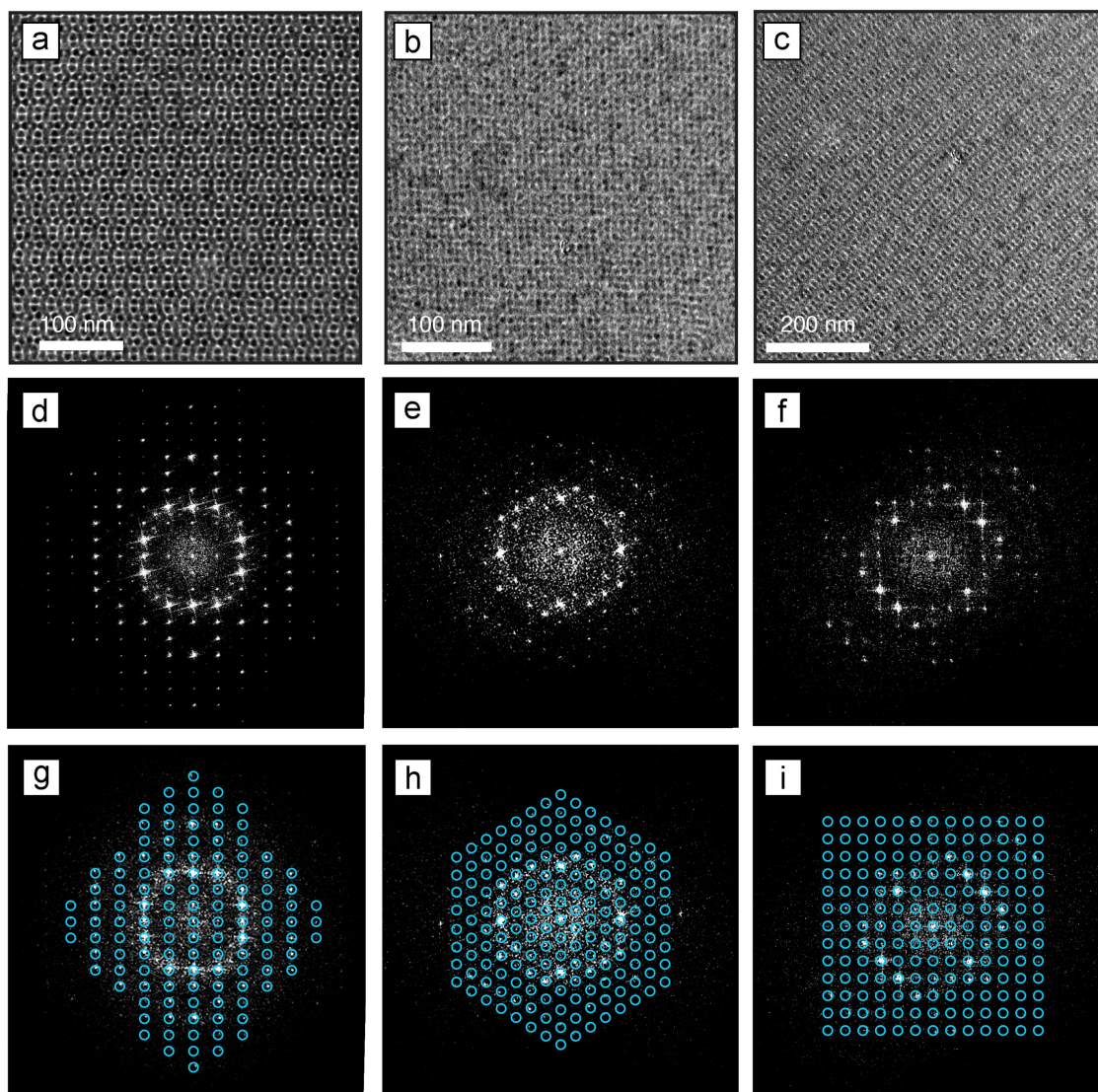
Extended Data Fig. 6 | Comparison between simulations and experimental results using characterization techniques in real and reciprocal spaces. **a**, TEM image for the monolayer superlattice of a bcc supercrystal along the $[\bar{1}10]_{\text{bcc}}$ projection. **b**, The FFT pattern of the image in **a**. **c**, SAXS method (ii) simulation of a bcc supercrystal from the $[\bar{1}10]_{\text{bcc}}$ projection (shown in Supplementary Figs. 19 and 20b). **d**, SAED pattern from the superlattice area of **a**. **e**, SAXS pattern of a bcc supercrystal

from the $[\bar{1}10]_{\text{bcc}}$ projection. **f**, WAED pattern from the superlattice area of **a**. **g**, WAXS pattern from a bcc supercrystal from the $[\bar{1}10]_{\text{bcc}}$ projection. We note that **b–e** all exhibited similar patterns in terms of the signals and the intensity profiles, indicating the validity of our approach to the identification of complex structures by iterating real and reciprocal spaces using the FFT of the model and the TEM images, diffraction patterns such as SAED and SAXS, and TEM images.



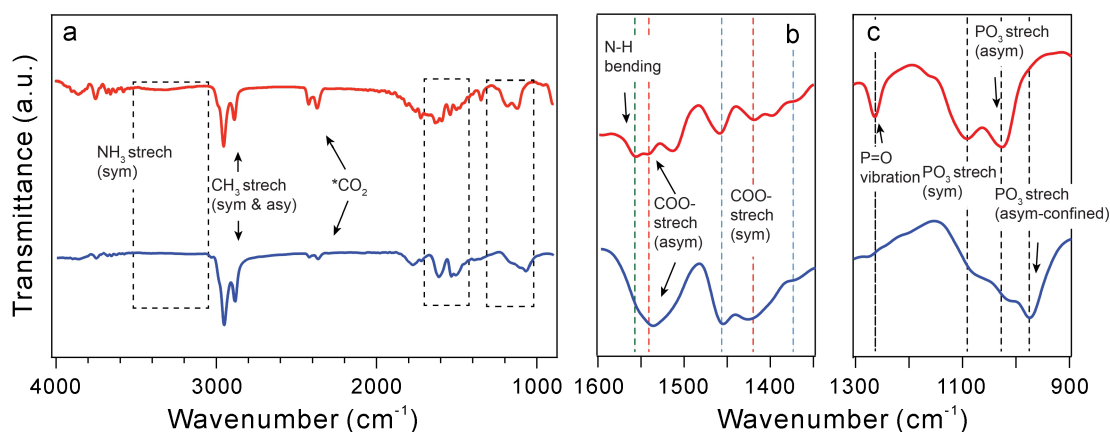
Extended Data Fig. 7 | TEM and SAED of bcc-supercrystal solids with tri-line contrast. a–h. Additional TEM images for bcc-supercrystal solids (a, c, e, g) and the corresponding SAED patterns (b, d, f, h). We note that although the TEM images all showed similar patterns with tri-line contrast (a, c, e, g), the SAED patterns for the selected areas gave different patterns (b, d, f, h). This discrepancy is due to a deeper penetration depth of the electron beam used in the SAED measurement than for the 2D

TEM imaging. This result indicates that these bcc supercrystals have a three-layer structure, and that the out-of-plane direction is parallel to the $[110]_{\text{bcc}}$ direction. We also assigned SAED patterns to the corresponding projections: the $[111]_{\text{bcc}}$ projection (b) and the $[100]_{\text{bcc}}$ projection (d). The SAED patterns shown here were obtained from a bcc supercrystal without fixed rotational orientations; therefore, some of the SAED patterns did not have clear bcc-oriented patterns such as those in f and h.



Extended Data Fig. 8 | TEM tilting experiment with monolayer superlattices of a bcc-supercrystal solid. **a–c**, TEM images from $[110]_{\text{bcc}}$ (**a**), $[111]_{\text{bcc}}$ (**b**) and $[010]_{\text{bcc}}$ (**c**) projections. **d–i**, The

corresponding FFT patterns from the TEM images shown in **a–c**. These patterns were assigned to the $[110]_{\text{bcc}}$ (**d**, **g**), $[111]_{\text{bcc}}$ (**e**, **h**) and $[010]_{\text{bcc}}$ (**f**, **i**) projections.



Extended Data Fig. 9 | FTIR spectra of an amorphous powder sample of the TTQDs and the bcc-supercrystal solid. In all panels, the spectra of the amorphous powder are shown in red and the bcc-supercrystal solid in blue. **a**, Full FTIR spectra. **b**, An expansion in the region of $1,600\text{ cm}^{-1}$ to $1,350\text{ cm}^{-1}$, which shows various peaks: a characteristic N–H bend appearing at $1,559\text{ cm}^{-1}$ (green dotted line); asymmetric and symmetric stretching of the carboxylate group ($-\text{COO}^-$) appearing at $1,537\text{ cm}^{-1}$ and $1,430\text{ cm}^{-1}$ respectively (red dotted lines); a C–H scissoring bend at $1,458\text{ cm}^{-1}$ and $\text{CH}_3\text{--C--H}$ bending at $1,378\text{ cm}^{-1}$ (blue dotted lines). The peak separation between asymmetric and symmetric stretching of the carboxylate group ($-\text{COO}^-$) is 107 cm^{-1} , which suggests bidentate binding and chelate formation. **c**, An expansion in the region of $1,300\text{ cm}^{-1}$

to 900 cm^{-1} , the P–O stretching region of the FTIR spectrum of ODPA molecules. The following peaks are observed: P=O vibration; PO_3 symmetric stretches in the range of $1,000\text{ cm}^{-1}$ to $1,150\text{ cm}^{-1}$; PO_3 asymmetric stretches in the range of 950 cm^{-1} to $1,030\text{ cm}^{-1}$. The PO_3 asymmetric stretch of the bcc-supercrystal sample is blue-shifted by 51 cm^{-1} (from $1,029\text{ cm}^{-1}$ to 978 cm^{-1}) when compared with the amorphous TTQD powder sample. The P=O vibration was also considerably weakened in the bcc-supercrystal sample in comparison to the amorphous TTQD powder sample. Similar changes have been observed previously, for example in ref. ²⁸, indicating a higher packing density of ODPA molecules in a confined state as a result of sandwiching between two atomic planes (that is, two $\{0002\}_{\text{WZ}}$ crystal facets).

Extended Data Table 1 | FTIR peak assignments for TTQDs and the bcc-supercrystal solid

peak (cm ⁻¹)	intensity	assignment	ref	amorphous TTQDs (cm ⁻¹)	<i>bcc-supercrystal</i> (cm ⁻¹)
3005	Medium	C-H stretch in C=C-H	25,S7	3001	3001
2924	Strong	CH ₂ asymmetric stretch	25,S7	2919	2919
2854	Strong	CH ₂ symmetric stretch	25,S7	2850	2850
1460-1470	Strong	C-H scissoring bending	S8,9	1458	1458
1378	Medium	CH ₃ -C-H bending	25,S9	1378	1378
1556-1540	Strong	COO- asymmetric stretch	S7,8	1537	1537
1418-1404	Strong	COO- symmetric stretch	S7,8	1430	1430
3000-3300	Weak	N-H stretch	25,30,S10	weak	absent
1558	Weak	N-H bending	25,30,S10	weak	absent
1261	Strong	P=O vibration	28,S11-14	1260	Weak
1000-1150	Strong	PO ₃ symmetric stretch	28,S11-14	1080	1088
950-1030	Strong	PO ₃ asymmetric stretch	28,S11-14	1029	978

The spectra are shown in Extended Data Fig. 9.

Ice loss from the East Antarctic Ice Sheet during late Pleistocene interglacials

David J. Wilson^{1,2*}, Rachel A. Bertram^{1,2}, Emma F. Needham¹, Tina van de Flierdt^{1,2}, Kevin J. Welsh³, Robert M. McKay⁴, Anannya Mazumder³, Christina R. Riesselman^{5,6}, Francisco J. Jimenez-Espejo^{7,8} & Carlota Escutia⁸

Understanding ice sheet behaviour in the geological past is essential for evaluating the role of the cryosphere in the climate system and for projecting rates and magnitudes of sea level rise in future warming scenarios^{1–4}. Although both geological data^{5–7} and ice sheet models^{3,8} indicate that marine-based sectors of the East Antarctic Ice Sheet were unstable during Pliocene warm intervals, the ice sheet dynamics during late Pleistocene interglacial intervals are highly uncertain^{3,9,10}. Here we provide evidence from marine sedimentological and geochemical records for ice margin retreat or thinning in the vicinity of the Wilkes Subglacial Basin of East Antarctica during warm late Pleistocene interglacial intervals. The most extreme changes in sediment provenance, recording changes in the locus of glacial erosion, occurred during marine isotope stages 5, 9, and 11, when Antarctic air temperatures¹¹ were at least two degrees Celsius warmer than pre-industrial temperatures for 2,500 years or more. Hence, our study indicates a close link between extended Antarctic warmth and ice loss from the Wilkes Subglacial Basin, providing ice-proximal data to support a contribution to sea level from a reduced East Antarctic Ice Sheet during warm interglacial intervals. While the behaviour of other regions of the East Antarctic Ice Sheet remains to be assessed, it appears that modest future warming may be sufficient to cause ice loss from the Wilkes Subglacial Basin.

The growth and decay of ice sheets act as important controls on regional and global climate by influencing albedo, sea ice extent, atmospheric and ocean circulation, nutrient supply, and sea level. In particular, the behaviour of the polar ice sheets is a key uncertainty in predicting sea level rise during and beyond this century^{1,2}. Complete melting of the marine-based West Antarctic Ice Sheet (WAIS) would contribute 3–5 m to global mean sea level, whereas the East Antarctic Ice Sheet (EAIS) contains a sea level equivalent of approximately 53 m, of which about 19 m is also marine-based ice¹². Although such ice is susceptible to loss through marine ice sheet instability processes where the bed deepens inland¹³, modelling studies differ in their predictions for both past and future stability of ice sheets^{2–4,8–10,14} (Fig. 1). It is therefore crucial to use geological evidence from proximal to ice sheet margins^{5–7,15,16} to constrain ice sheet responses to climate forcing.

The late Pleistocene interval represents a critical target for exploring ice sheet behaviour because of similar ice sheet boundary conditions to the present day⁸ and well-constrained evidence on global mean sea levels¹⁷ and regional climatic forcing^{11,18,19}. Global mean sea levels were 6–9 m higher than at present during interglacial marine isotope stage (MIS) 5e, and 6–13 m higher than at present during interglacial MIS 11, probably requiring a substantial reduction in Antarctic ice¹⁷. Sedimentological evidence suggests that the Pleistocene ice sheets draining into the Ross Sea Embayment fluctuated on glacial–interglacial timescales¹⁵, and it has been proposed that the WAIS collapsed during at least one Pleistocene interglacial¹⁶. Although proximal marine sedimentary records indicate that both the WAIS and the EAIS had

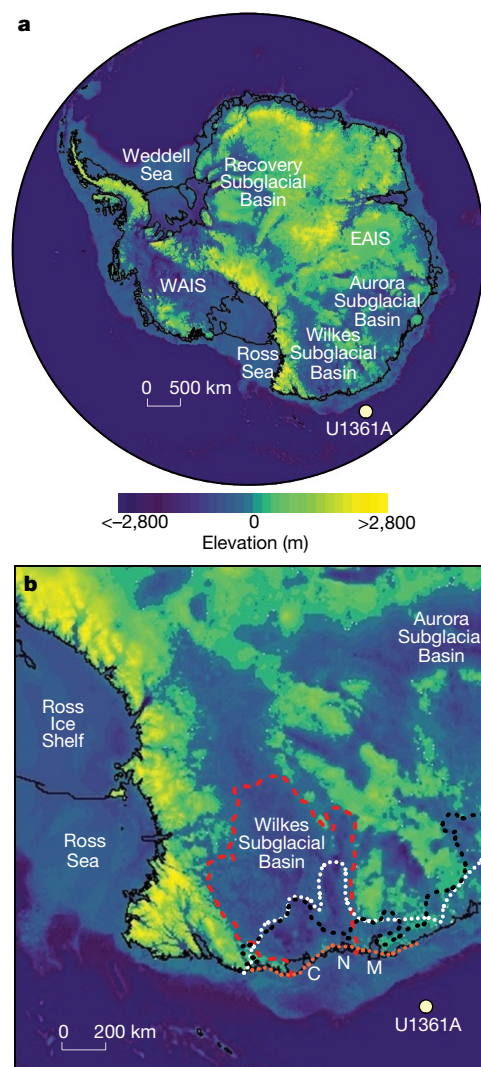


Fig. 1 | Setting of IODP Site U1361 offshore of the Wilkes Subglacial Basin. **a**, Map of Antarctica showing subglacial bedrock elevation above sea level^{12,31} and the U1361A coring location. **b**, Detailed map of the Wilkes Subglacial Basin, with lines illustrating positions of the ice sheet margin in different ice sheet models and scenarios: red dashed line, fully retreated state of Mengel and Levermann² under 1.8°C ocean warming; black dashed line, maximum simulated MIS 5e retreat of DeConto and Pollard³, equivalent to approximately 2°C ocean and atmospheric warming; and modelled retreat of Golledge et al.⁹ for both 2°C ocean and atmospheric warming (ochre dotted line) and 4°C ocean and atmospheric warming (white dotted line). C, N, and M indicate positions of Cook, Ninnis, and Mertz ice shelves, respectively.

¹Department of Earth Science and Engineering, Imperial College London, London, UK. ²Grantham Institute - Climate Change and the Environment, Imperial College London, London, UK. ³School of Earth and Environmental Sciences, University of Queensland, Brisbane, Queensland, Australia. ⁴Antarctic Research Centre, Victoria University of Wellington, Wellington, New Zealand. ⁵Department of Geology, University of Otago, Dunedin, New Zealand. ⁶Department of Marine Science, University of Otago, Dunedin, New Zealand. ⁷Department of Biogeochemistry, JAMSTEC, Yokosuka, Japan.

⁸Andalusian Institute of Earth Sciences, CSIC and Universidad de Granada, Armilla, Spain. *e-mail: david.wilson1@imperial.ac.uk

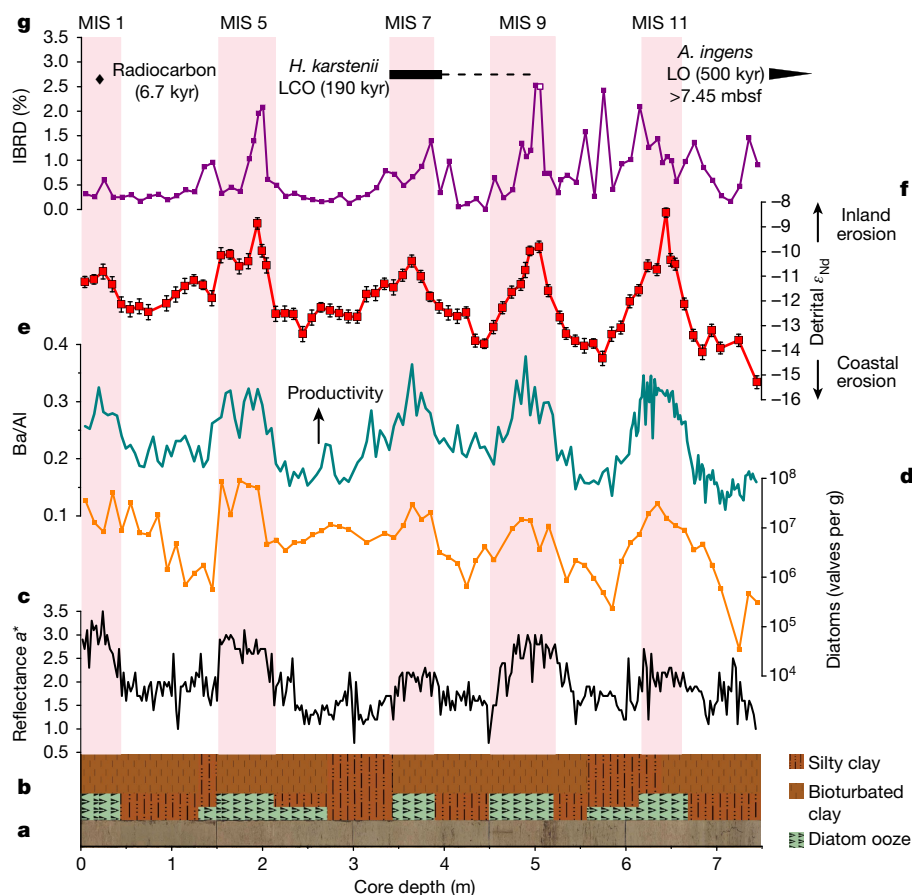


Fig. 2 | Late Pleistocene records from U1361A spanning MIS 1 to MIS 12. **a**, Core images. **b**, Lithological log. **c**, Sediment reflectance (a^*). **d**, Absolute diatom abundance (logarithmic scale). **e**, Ba/Al ratios (X-ray fluorescence (XRF)-scanner counts). **f**, Detrital sediment Nd isotopes (arrows indicate coastal erosion of Lower Palaeozoic granitoids versus inland erosion of FLIP/Beacon lithologies; error bars are 2 s.d. external reproducibility). **g**, Ice berg rafted debris (per cent 250 μm –2 mm); open square represents

one data point that plots off the scale with 7.5% IBRD at 5.05 m depth). Pink columns highlight interglacial periods, based on the dominance of diatom-rich clay. MIS numbers are labelled at the top, together with chronostratigraphic constraints (LCO, last common occurrence; LO, last occurrence; see Methods). Data in **a**–**c** from ref. ²⁰; all other data from this study (Supplementary Tables 1, 4, 5, 7, 8). mbsf, metres below sea floor.

dynamic margins during warm Pliocene intervals^{5–7,15}, there are currently no data that directly constrain EAIS behaviour during the late Pleistocene.

Here we provide new observations that constrain the behaviour of a marine-based sector of the EAIS during the late Pleistocene, based on sedimentological and geochemical records in marine sediment core U1361A (64.41°S, 143.89°E, 3,454 m water depth), recovered from the continental rise offshore of the Wilkes Subglacial Basin (Fig. 1) during the Integrated Ocean Drilling Program (IODP) Expedition 318²⁰. Because it has a landward-sloping bed and 3–4 m sea level equivalent contained in marine-based ice^{2,12}, the Wilkes Subglacial Basin represents both a sensitive test for EAIS vulnerability⁹ and a substantial potential contributor to global sea level. Glacial–interglacial cycles are well resolved in the upper 7.5 m of U1361A, with glacial intervals indicated by diatom-poor silty clays with occasional laminations and low barium/aluminium (Ba/Al) ratios, and interglacial intervals marked by bioturbated diatom-rich clays with high Ba/Al ratios, high sediment reflectance and occasional dropstones²⁰ (Fig. 2a–e). These lithological cycles record changes in local productivity that are likely to have resulted from reduced presence of perennial sea ice during interglacials²⁰. Assignment of the diatom-rich intervals to the past five interglacials (pink bars in Fig. 2) is well supported by all available chronostratigraphic data (see Methods; Extended Data Fig. 3; Supplementary Table 8).

As sediment supply to the continental rise at Site U1361 is dominated by downslope transport from the proximal shelf²⁰ (see Methods), sediment provenance in this location is sensitive to processes at the

regional ice sheet margin. To constrain provenance changes through time, we analysed neodymium (Nd) isotopic compositions on the bulk sediment fraction (see Methods) at approximately 10-cm intervals (6,000 year resolution). These data are expressed as ϵ_{Nd} , the deviation of $^{143}\text{Nd}/^{144}\text{Nd}$ ratios from the Chondritic Uniform Reservoir value in parts per 10,000. Because Nd isotopes vary as a function of the age and lithology of the eroded source rocks, they are an appropriate tracer for changes in the distribution of glacial erosion in the vicinity of the Wilkes Subglacial Basin⁵. Sediment grain size (per cent 250 μm –2 mm) was also measured to provide complementary evidence on the supply of ice-berg-rafted debris (IBRD) from calving ice margins (see Methods).

Our detrital Nd isotope record reveals variability between $\epsilon_{\text{Nd}} = -15.5$ to -12 during glacials and $\epsilon_{\text{Nd}} = -12$ to -8.5 during interglacials (Fig. 2f), corresponding closely with sea-ice retreat and productivity changes inferred from Ba/Al ratios (Fig. 2e; see Methods). Provenance shifts towards more radiogenic Nd isotopic compositions at the onset of interglacials MIS 5 and MIS 9 coincide closely with peaks in the supply of IBRD, whilst IBRD peaks slightly precede the provenance changes for MIS 7 and MIS 11 (Fig. 2f, g). Whereas IBRD peaks are transient events that record dynamic ice discharge, typically during deglaciation⁶ (see Methods), the provenance changes are sustained for longer and suggest a prolonged switch in the locus of regional glacial erosion. Additional minor IBRD peaks occur near the ends of diatom-rich intervals, which might record an advancing ice margin at the onset of glaciation.

The sediment provenance changes at U1361A during the late Pleistocene are strikingly similar in pattern and magnitude to changes

reported previously for the Pliocene⁵, which ranged from $\varepsilon_{\text{Nd}} = -15$ to -6 . Hence, we infer that similar lithologies were eroded during both periods, and this is further supported by detrital strontium (Sr) isotope data (Extended Data Fig. 2). In detail, unradiogenic Nd isotopic compositions during late Pleistocene glacials ($\varepsilon_{\text{Nd}} = -15.5$ to -12) record a dominant Lower Palaeozoic provenance ($\varepsilon_{\text{Nd}} = -20$ to -10), which is explained by bedrock erosion of Lower Palaeozoic granitoids near the modern ice margin, particularly around the Ninnis Glacier⁵ (Extended Data Fig. 1). By contrast, the more radiogenic compositions during interglacials (up to ε_{Nd} values of -8.5) require additional contributions from a more radiogenic source that is inferred to be the Permian–Cretaceous rocks of the Ferrar Large Igneous Province (FLIP)⁵ ($\varepsilon_{\text{Nd}} = -7$ to -3.5). Such FLIP basalts and dolerites, and associated Beacon Supergroup sediments, are not exposed in substantial areas at the present ice margin, with only a small coastal outcrop at Horn Bluff (Extended Data Fig. 1), but are inferred to be extensively present within the ice-covered Wilkes Subglacial Basin^{5,21}. The admixture of a FLIP component within the interglacial sediments at U1361A is independently supported by acid-reductive leaching experiments, which extract a reactive component with a Nd isotopic composition consistent with the FLIP (Extended Data Fig. 1; see Methods).

An enhanced contribution of FLIP lithologies to Site U1361 during warmer than present interglacials (Fig. 2f) must indicate a shift in the erosional regime of the ice sheet and/or a change in sediment transport to the site. Although the flow speed of the wind-driven westward-flowing Antarctic Coastal Current²² could have changed through time, we argue against sediment transport changes as the dominant control on our provenance record. First, warming is proposed to weaken the polar easterlies and reduce the strength of the Antarctic Coastal Current^{22,23}, whereas an enhanced supply of FLIP-derived sediments from upstream coastal sources (for example, Horn Bluff) during warm interglacials would instead require a strengthened current. Second, we observe similar Nd isotope values and temporal variability in the bulk and fine (less than $63\ \mu\text{m}$) sediment fractions for ten co-analysed samples in U1361A (Extended Data Fig. 1; see Methods), which rules out mineralogical or grain size sorting as a major control on the record. Third, a record of sortable silt mean grain size in U1361A (Supplementary Table 4) provides no evidence for substantial deep flow speed changes, and shows no co-variation with detrital Nd isotopes, which is consistent with a turbidite-dominated transport environment but inconsistent with deep current transport driving the radiogenic Nd isotope signal (see Methods).

Because ice sheets predominantly erode bedrock near their margins²⁴, we consider past interglacial retreat of the ice margin into the Wilkes Subglacial Basin^{2,3,9} (Fig. 1b), coupled to enhanced erosion of FLIP bedrock and/or FLIP-derived sediment infill of the Central Basin²¹, to be the most likely explanation for the provenance changes at Site U1361. In addition, enhanced erosion of FLIP and associated Beacon lithologies close to the mouth of the Wilkes Subglacial Basin could have contributed to the provenance shift. Without necessitating full-scale retreat, this latter scenario would require a substantial change in the erosion distribution, implicating flow acceleration linked to ice sheet thinning²⁴. Although we cannot quantitatively determine the relative importance of margin retreat versus thinning, we emphasize that either scenario would imply a close link between detrital provenance and ice volume reduction in the Wilkes Subglacial Basin, and hence a late Pleistocene sea level contribution from the EAIS.

The key finding from our new data set is that the Wilkes Subglacial Basin has been susceptible to ice loss not only during warm Pliocene intervals⁵ with CO_2 levels of approximately 400 p.p.m., but also during the late Pleistocene despite CO_2 levels²⁵ remaining below 300 p.p.m. Hence, we provide data-based evidence in support of recent ice sheet models that simulate margin retreat and ice loss during late Pleistocene interglacials^{2,3,9} (Fig. 1b). Although global climate forcing during recent interglacials was generally similar to the Holocene, evidence from both Antarctica¹¹ and the Southern Ocean^{18,19} reveals that a number of these interglacials were at times characterized by warmer temperatures than the Holocene (Fig. 3a–c), consistent with a link between regional

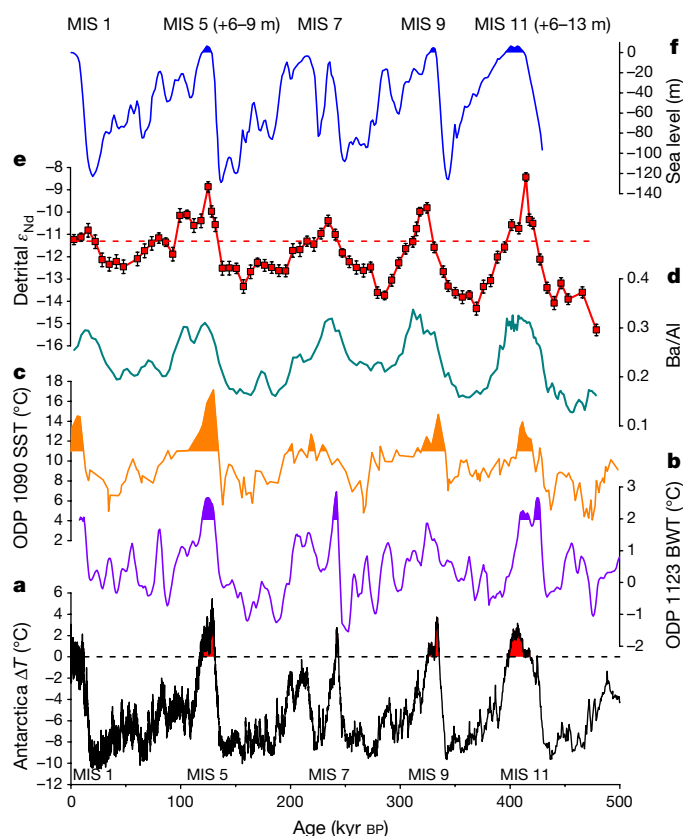


Fig. 3 | Comparison of U1361A records to regional palaeoclimate and global sea level records. a, Antarctic ice core temperature difference (ΔT , difference from mean values of the last millennium) derived from deuterium isotopes at EPICA Dome C (EDC)¹¹ plotted on EDC3 age scale. **b**, before present. **b**, Southern Ocean bottom water temperature (BWT) from Mg/Ca at Ocean Drilling Program (ODP) Site 1123 (ref. 18). **c**, Southern Ocean sea surface temperature (SST) from alkenones at ODP Site 1090 (ref. 19). **d**, Ba/Al ratios (XRF-scanner counts; three-point smoothed) in U1361A. **e**, Bulk detrital sediment Nd isotopes in U1361A (error bars are 2 s.d. external reproducibility). **f**, Sea level proxy from benthic oxygen isotopes²⁸, labelled with MIS numbers and sea level estimates¹⁷ from MIS 5e and MIS 11. Shading in **a–c**, **f** represents intervals with values above modern (or late Holocene core top); red dashed line in **e** indicates the core top ε_{Nd} value of U1361A. For chronostratigraphic constraints on U1361A, see Supplementary Table 8 and Methods.

warming and ice sheet retreat^{2,3}. Possible causes of transient warmth include a bipolar seesaw tied to variability of the Atlantic meridional overturning circulation¹¹, wind shifts that enhanced Circumpolar Deep Water upwelling onto the shelves²⁶, or feedback effects linked to the size of the WAIS²⁷.

In detail, it is striking that differences in the peak Nd isotopic compositions of individual interglacials correlate well with Southern Ocean temperatures^{11,18,19} (Fig. 3a–c) and global sea level^{17,28} (Fig. 3f). In particular, MIS 5, 9 and 11, with the most distinct radiogenic Nd isotope excursions (Fig. 3e), experienced the most extended warmth in Antarctica¹¹ (Fig. 3a), with temperatures more than 2°C warmer than the pre-industrial late Holocene for approximately 4,000 years, 2,500 years and 6,000 years, respectively. By contrast, MIS 7 had a more muted Nd isotope response, comparable to the transition from the Last Glacial Maximum into the Holocene (Fig. 3e), and experienced temperatures more than 2°C warmer than the pre-industrial late Holocene for only approximately 1,000 years (Fig. 3a). This scaling between the magnitude and duration of regional climate warming, the Nd isotope-based provenance variations, and global sea level (Fig. 3) is consistent with an ice volume change in the Wilkes Subglacial Basin. We therefore suggest that an Antarctic warming of about 2°C above pre-industrial temperatures for approximately 1,000–2,500 years has been sufficient to cause

ice loss beyond that of the modern day or pre-industrial Holocene, leading to a contribution to global sea levels from the EAIS during MIS 5, 9 and 11, in agreement with recent modelling results³ for MIS 5e.

Differences in provenance between individual interglacials (Fig. 3e) also hint at the underlying mechanisms of ice retreat in the Wilkes Subglacial Basin, by supporting variable ice margin retreat during climate warming, rather than a simple switch between advanced and retreated states² (Fig. 1b). Variable retreat could reflect the presence of multiple pinning points on basement highs^{9,24} (Fig. 1b), which represent temporary or final limits for the ice margin. In this view, the more extreme radiogenic Nd isotope excursions observed during some (but not all) warm Pliocene intervals⁵ may indicate more substantial ice sheet retreat at these times.

Although we have focused on ice sheet behaviour during individual interglacials, our data also provide an indication of longer-timescale changes in ice sheet dynamics. Over the last 480,000 years, there has been a long-term trend towards declining IBRD (Fig. 2g), suggesting an increasingly stable ice margin, while glacial ϵ_{Nd} values have become less extreme (that is, more radiogenic; Fig. 2f), potentially indicating reduced glacial advance over the shelf. Together, these observations appear consistent with gradual stabilization of the ice margin as a consequence of grounding zone sediment accumulation through multiple glacial cycles^{29,30}, but this hypothesis awaits further testing.

By providing geological evidence for ice loss from a dynamic margin of the EAIS during recent warm interglacial intervals (MIS 5, 9 and 11), our data place new constraints on the role of Antarctica in past sea-level changes, and represent a useful target for ice sheet models. Based on the ice sheet response during past interglacial periods, we estimate that substantial ice loss within the Wilkes Subglacial Basin would be likely to occur with approximately 2 °C warming (above pre-industrial) if sustained for a few millennia. This scenario is broadly consistent with the magnitudes and timescales of forcing that generate ice margin retreat in models^{2–4}. While evidence from sediment provenance cannot precisely quantify the magnitude of any sea level contribution, our data appear to suggest that some models (for example, refs. ^{9,14}) may have underestimated the long-term potential of the Wilkes Subglacial Basin, and perhaps the EAIS more generally, to contribute to future sea level rise.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0501-8>.

Received: 16 November 2017; Accepted: 25 July 2018;

Published online 19 September 2018.

1. IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds. Stocker, T. F. et al.) 1535 (Cambridge Univ. Press, Cambridge, 2013).
2. Mengel, M. & Levermann, A. Ice plug prevents irreversible discharge from East Antarctica. *Nat. Clim. Chang.* **4**, 451–455 (2014).
3. DeConto, R. M. & Pollard, D. Contribution of Antarctica to past and future sea-level rise. *Nature* **531**, 591–597 (2016).
4. Golledge, N. R. et al. The multi-millennial Antarctic commitment to future sea-level rise. *Nature* **526**, 421–425 (2015).
5. Cook, C. P. et al. Dynamic behaviour of the East Antarctic ice sheet during Pliocene warmth. *Nat. Geosci.* **6**, 765–769 (2013).
6. Patterson, M. O. et al. Orbital forcing of the East Antarctic ice sheet during the Pliocene and Early Pleistocene. *Nat. Geosci.* **7**, 841–847 (2014).
7. Reinardy, B. T. I. et al. Repeated advance and retreat of the East Antarctic Ice Sheet on the continental shelf during the early Pliocene warm period. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **422**, 65–84 (2015).
8. Austermann, J. et al. The impact of dynamic topography change on Antarctic ice sheet stability during the mid-Pliocene warm period. *Geology* **43**, 927–930 (2015).
9. Golledge, N. R., Levy, R. H., McKay, R. M. & Naish, T. R. East Antarctic ice sheet most vulnerable to Weddell Sea warming. *Geophys. Res. Lett.* **44**, 2343–2351 (2017).
10. Pollard, D. & DeConto, R. M. Modelling West Antarctic ice sheet growth and collapse through the past five million years. *Nature* **458**, 329–332 (2009).
11. Jouzel, J. et al. Orbital and millennial Antarctic climate variability over the past 800,000 years. *Science* **317**, 793–796 (2007).

12. Fretwell, P. et al. Bedmap2: improved ice bed, surface and thickness datasets for Antarctica. *Cryosphere* **7**, 375–393 (2013).
13. Schoof, C. Ice sheet grounding line dynamics: Steady states, stability, and hysteresis. *J. Geophys. Res. Earth Surf.* **112**, (2007).
14. Ritz, C. et al. Potential sea-level rise from Antarctic ice-sheet instability constrained by observations. *Nature* **528**, 115–118 (2015).
15. Naish, T. et al. Obliquity-paced Pliocene West Antarctic ice sheet oscillations. *Nature* **458**, 322–328 (2009).
16. Scherer, R. P. et al. Pleistocene collapse of the West Antarctic ice sheet. *Science* **281**, 82–85 (1998).
17. Dutton, A. et al. Sea-level rise due to polar ice-sheet mass loss during past warm periods. *Science* **349**, aaa4019 (2015).
18. Elderfield, H. et al. Evolution of ocean temperature and ice volume through the Mid-Pleistocene Climate Transition. *Science* **337**, 704–709 (2012).
19. Martínez-García, A. et al. Links between iron supply, marine productivity, sea surface temperature, and CO₂ over the last 1.1 Ma. *Paleoceanography* **24**, PA1207 (2009).
20. Escutia, C., Brinkhuis, H., Klaus, A. & the Expedition 318 Scientists *Proc. IODP, 318* (Integrated Ocean Drilling Program Management International, Inc., Tokyo, 2011).
21. Ferraccioli, F., Armadillo, E., Jordan, T., Bozzo, E. & Corr, H. Aeromagnetic exploration over the East Antarctic Ice Sheet: A new view of the Wilkes Subglacial Basin. *Tectonophysics* **478**, 62–77 (2009).
22. Spence, P. et al. Rapid subsurface warming and circulation changes of Antarctic coastal waters by poleward shifting winds. *Geophys. Res. Lett.* **41**, 4601–4610 (2014).
23. DeConto, R., Pollard, D. & Harwood, D. Sea ice feedback and Cenozoic evolution of Antarctic climate and ice sheets. *Paleoceanography* **22**, PA3214 (2007).
24. Golledge, N. R. et al. Antarctic climate and ice-sheet configuration during the early Pliocene interglacial at 4.23 Ma. *Clim. Past* **13**, 959–975 (2017).
25. Siegenthaler, U. et al. Stable carbon cycle-climate relationship during the late Pleistocene. *Science* **310**, 1313–1317 (2005).
26. Fogwill, C. J. et al. Testing the sensitivity of the East Antarctic Ice Sheet to Southern Ocean dynamics: past changes and future implications. *J. Quat. Sci.* **29**, 91–98 (2014).
27. Holden, P. B. et al. Interhemispheric coupling, the West Antarctic Ice Sheet and warm Antarctic interglacials. *Clim. Past* **6**, 431–443 (2010).
28. Waelbroeck, C. et al. Sea-level and deep water temperature changes derived from benthic foraminifera isotopic records. *Quat. Sci. Rev.* **21**, 295–305 (2002).
29. Alley, R. B., Anandakrishnan, S., Dupont, T. K., Parizek, B. R. & Pollard, D. Effect of sedimentation on ice-sheet grounding-line stability. *Science* **315**, 1838–1841 (2007).
30. Pollard, D. & DeConto, R. M. in *Glacial Sedimentary Processes and Products* (eds. Hambrey, M.J., Christoffersen, P., Glasser, N.F. & Hubbard, B.) 37–52 (Blackwell, Oxford, 2007).
31. Thompson, J. W. & Cooper, A. P. R. The SCAR Antarctic digital topographic database. *Antarctic Sci.* **5**, 239–244 (1993).

Acknowledgements This research used samples and data provided by Integrated Ocean Drilling Program (IODP) Expedition 318, sponsored by the US National Science Foundation (NSF) and participating countries under the management of the Consortium for Ocean Leadership. D.J.W. thanks B. Coles, C. Huck, K. Kreissig, N. Pratt and P. Simoes Pereira for technical support. D.J.W., R.A.B., E.F.N. and T.v.d.F. acknowledge financial support from the Kristian Gerhard Jebsen Foundation, the Leverhulme Trust (RPG-398) and NERC (NE/N001141/1, NE/H025162/1). K.J.W. and R.M.M. were funded by the Australia-New Zealand IODP Consortium's Australian Research Council LIEF grants (LE140100047, LE0882854). R.M.M. was funded by a Royal Society (New Zealand) Rutherford Discovery Fellowship (RDF-13-VUW-003). C.E. and F.J.J.-E. acknowledge funding from the Spanish Ministry of Science and Innovation Grant CTM2017-89711-C2-1 co-financed by the European Regional Development Fund (FEDER).

Reviewer information *Nature* thanks A. Shevenell and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.J.W., T.v.d.F. and K.J.W. designed the research; D.J.W., R.A.B., E.F.N., and T.v.d.F. carried out the Nd isotope analyses; R.A.B. carried out the Sr isotope analyses; A.M. performed the diatom counts with guidance from C.R.R. and K.J.W.; R.M.M. and K.J.W. carried out sedimentological analyses; F.J.J.-E. and C.E. conducted XRF scanning measurements and PCA analysis; C.R.R., K.J.W., and R.M.M. generated the age model. All authors contributed to data interpretation. D.J.W. wrote the paper, with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0501-8>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0501-8>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to D.J.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Detrital sediment Nd isotope measurements. Seventy-five samples from U1361A were analysed for detrital sediment Nd isotopes (Supplementary Table 1), after leaching to remove any authigenic ferromanganese oxide phases^{5,32–34}. Bulk samples of ~0.5–1 g were leached twice in an acid-reductive solution of hydroxylamine hydrochloride/acetic acid, including a second leach of at least 12 h in 0.02 M hydroxylamine hydrochloride/25% buffered acetic acid. After crushing and homogenization, digestion of the detrital residue was carried out on a ~50 mg subsample by hotplate digestion⁵ using a mixture of 2 ml 23 M HF, 1 ml 16 M HNO₃, and 0.6 ml 20 M HClO₄ (repeated twice), and then a mixture of 2 ml 23 M HF and 1 ml 16 M HNO₃. The rare earth element fraction was separated using either Eichrom TRU spec resin (100–120 µm mesh) or cation exchange resin (200–400 µm mesh), and the Nd fraction was isolated using Eichrom LN spec resin (50–100 µm mesh) on volumetrically calibrated Teflon columns⁵. Replacement of TRU spec resin by cation exchange resin was found to improve column yields.

Neodymium isotopic compositions were analysed on the Nu Plasma multi-collector inductively coupled plasma mass spectrometer (MC-ICP-MS) in the MAGIC laboratories at Imperial College London. Mass bias was corrected using the exponential law (to ¹⁴⁶Nd/¹⁴⁴Nd = 0.7219). A correction for direct ¹⁴⁴Sm interference was also applied, with all samples significantly below the threshold determined for accurate correction (<0.1% of the ¹⁴⁴Nd signal). Sample measurements were bracketed by concentration-matched JNdi-1 Nd isotope standards, and data from each analytical session were adjusted to give agreement with the literature value³⁵. The external reproducibility was estimated from the within-session standard deviation (2 s.d.) on those standards. Over the course of analyses, measurements of rock standard BCR-2 gave ¹⁴³Nd/¹⁴⁴Nd = 0.512640 ± 0.000016 (*n* = 31), in excellent agreement with literature values³⁶, and indicating a long-term reproducibility of 0.31 ε_{Nd} units. Procedural and column blanks were typically 4–18 pg, and hence negligible.

Robustness of the detrital sediment Nd isotope record. Unlike some other radiogenic isotope systems (for example, Sr and Hf), Nd isotopes are relatively insensitive to grain size fractionation³⁷, making them a robust indicator of sediment provenance. However, if changes in sediment transport processes led to the supply of sediment from different sources, a link between grain size and Nd isotopes could emerge, which would complicate provenance interpretations. To test for any potential bias in our detrital Nd isotope record measured on the bulk sediment fraction, we additionally measured Nd isotopes on the fine-grained fraction (<63 µm) for a subset of 10 samples, including both peak glacial and interglacial conditions (Supplementary Table 2). The fine fraction was separated by wet sieving and then subject to leaching, digestion, Nd separation and isotope measurement identically to the bulk sediment. Procedural blanks were 1.3 ng, representing less than 0.02% of sample Nd and hence negligible, while all five full procedural replicates yielded Nd isotope results that agreed within error (Supplementary Table 2).

Since comparable glacial–interglacial Nd isotope changes are recorded in the <63 µm fraction and the bulk fraction (Extended Data Fig. 1), it is clear that grain size (potentially linked to sediment transport) has an insignificant influence on the detrital Nd isotope record. In detail, two out of ten samples (from interglacials MIS 5 and MIS 11) show a small but resolvable offset of around 1 ε_{Nd} unit towards more radiogenic values in the bulk sediment than the fine fraction (Extended Data Fig. 1). This observation further constrains that the radiogenic Nd isotope signal from FLIP/Beacon lithologies is not restricted to a fine-grained silt or clay fraction but is also contained within sand-sized material that could be derived from turbidity currents or ice-rafting. Given the potential of ocean currents to transport material only within the silt fraction³⁸, this evidence is consistent with a local source rather than a strengthened alongslope current in driving the radiogenic Nd isotope excursions.

A final indication that grain size and/or sediment transport are not major controls on the detrital Nd isotope record is given by the lack of a direct correlation between the IBRD record (per cent 250 µm–2 mm fraction) and detrital Nd isotopes (Fig. 2). While there were generally IBRD peaks during deglaciation, the Nd isotope changes were often sustained throughout interglacials while IBRD was generally minimal or absent. Hence, the mechanism of sediment delivery appears to have been decoupled from the provenance information, which supports the complementary interpretation of IBRD as an indicator of dynamic processes (see below) and Nd isotopes as an indicator of the zone of glacial erosion. As previously shown in this region³⁹, provenance information from the bulk or fine sediment fraction is more useful as an indicator of glacial erosion than evidence from the IBRD fraction, hence our focus here on the former approach.

Supporting evidence on provenance changes from reductive sediment leachates. Acid-reductive sediment leaching is used to extract authigenic ferromanganese oxides or other reactive phases that are dispersed within a sediment core^{32–34,40}. In many settings, this approach is useful for reconstructing past deep seawater compositions, but in certain cases it has been shown to yield compositions that reflect local sedimentary inputs, either through their partial dissolution during

extraction³² or through in situ exchange with pore water⁴¹. Such a local influence is most likely at ocean margin sites where terrigenous-rich sediments may contain preformed ferromanganese oxides⁴², or in locations where reactive volcanics are present in the sediment^{32,41,43,44}.

In this study, we analysed acid-reductive sediment leachates, based on two published methods^{32,40}, including 1 h leaches (*n* = 18) and subsequent overnight leaches (*n* = 6) (Supplementary Table 3). During glacial periods, the leachate Nd isotopic compositions (ε_{Nd} = −5.6 to −10.5; Extended Data Fig. 1) are consistent with the range of possible compositions of Circumpolar Deep Water or Antarctic Bottom Water^{45,46}, although an influence of a mixed detrital component (Lower Palaeozoic and FLIP/Beacon) on the recorded signature cannot be ruled out. However, during interglacial periods (particularly MIS 5, 9, and 11), the leachate compositions (ε_{Nd} = −3.2 to −6.1) are too radiogenic to reflect the composition of any likely local bottom water, but are instead consistent with a major influence from a reactive basaltic FLIP source (Extended Data Fig. 1). Therefore, the leachate data provide strong independent support for our provenance interpretations based on the detrital sediment Nd isotope record.

Detrital sediment Sr isotope measurements. Measurements of Sr isotopes were made on the same subset of 10 fine-grained sediment fractions (<63 µm) that were analysed for Nd isotopes, spanning both peak glacial and interglacial conditions (Supplementary Table 2). Following rare earth element separation, samples were processed through Eichrom Sr-Spec resin (100–120 µm mesh) to isolate the Sr fraction. Strontium isotopes were analysed on the Thermo Scientific Triton thermal ionisation mass spectrometer (TIMS) in the MAGIC laboratories at Imperial College London. Samples were loaded onto degassed single tungsten filaments in 1 µl 6 M HCl, followed by 1 µl of tantalum chloride activator. Measurements were made in static mode, with instrumental mass bias corrected using the exponential law (to ⁸⁸Sr/⁸⁶Sr = 8.375), and interferences of ⁸⁷Rb corrected using an ⁸⁷Rb/⁸⁵Rb ratio of 0.3860. Long-term repeated analysis of the NBS 987 Sr standard in the MAGIC laboratories yielded ⁸⁷Sr/⁸⁶Sr = 0.710247 ± 0.000019 (2 s.d., *n* = 84) and reported ⁸⁷Sr/⁸⁶Sr ratios were corrected to the published value³⁶ (0.710252 ± 0.000013). Ten separate digests of rock standard BCR-2 gave ⁸⁷Sr/⁸⁶Sr = 0.705010 ± 0.000014 (2 s.d., *n* = 42), in good agreement with the published value³⁶ (0.705013 ± 0.000010). The procedural blank associated with this sample set was ~6 ng, representing less than a 0.5% contribution to sample measurements.

Stability of provenance endmembers from combined Nd and Sr isotopes. The combined Nd and Sr isotope data set for the late Pleistocene is fully consistent with mixing between Lower Palaeozoic and FLIP/Beacon lithologies, and also records an identical trend to Pliocene data from U1361 (Extended Data Fig. 2). We therefore infer similar provenance variations for both intervals, which additionally appears to indicate that any long term erosional changes⁴⁷ have not affected the regional bedrock sources available for erosion over this period of time.

IBRD measurements. Ninety-seven samples from U1361A were processed for grain size analysis at Victoria University of Wellington to provide evidence on IBRD content (Supplementary Table 4). Dried samples were weighed and then wet-sieved to recover the coarse sand fraction (250 µm to 2 mm), which has previously been used to indicate IBRD in Arctic and Antarctic studies^{6,48}. The >250 µm fraction was dissolved of biogenic silica using 2M NaOH, and then dried and weighed to obtain a weight per cent of IBRD fraction (250 µm–2 mm) relative to the bulk sediment. Each sample was visually examined again under binocular microscope for volcanic ash layers, as well as for any authigenic minerals or biogenic components that could bias the IBRD weights, and such components were manually removed if present. For this study, we did not calculate IBRD mass accumulation rates, but they would scale almost directly with the weight percent IBRD (since our age model assumes constant sedimentation rates and there are no large shifts or trends in dry bulk density in this interval).

Interpretation of IBRD record. Antarctica currently loses approximately half of its ice mass via iceberg calving⁴⁹, while distinct pulses of IBRD observed in the Scotia Sea region during the last deglaciation have been interpreted as representing collapse of the marine-based sector of the WAIS⁵⁰. Site U1361 is located in the pathway of the Antarctic Coastal Current which transports icebergs into the region in a predictable manner, as opposed to regions further north or east⁵¹. Furthermore, it is located south of the present southern boundary of the Antarctic Circumpolar Current (ACC), in a region dominated by geostrophic and bathymetrically controlled currents. Since the southern boundary front of the ACC is bathymetrically controlled by the continental rise on which Site U1361 is located, changes in ocean currents or sea surface temperature are unlikely to have contributed significantly to changes recorded in our IBRD record. Site U1361 is also a suitable site for this proxy because it is proximal to outlet glaciers of the Antarctic margin, such that smaller icebergs with basal debris could survive moderate levels of sea surface warming (even to the elevated temperatures inferred for the Pliocene^{5,6}), but also not so close to the continent as to be influenced by a single outlet glacier or a single iceberg dumping event⁵¹.

Studies from the Pliocene interval of U1361 indicate that the IBRD record contains a statistically significant, high-fidelity orbital signal, confirming that iceberg calving is not a random process at this site⁶. As in that study, we interpret the new IBRD record (Fig. 2) to primarily represent changes in the dynamic discharge of outlet glaciers and ice streams, likely occurring during periods of enhanced glacial retreat. One potential complicating factor in the interpretation of this record is that the development of large ice shelves during glacial maxima and minima could theoretically result in tabular icebergs that lack basal debris, such as for the modern Mertz Glacier tongue or the Ross Ice Shelf. In this scenario, our IBRD record (Fig. 2) may document only the calving of smaller icebergs from outlet glaciers, such that the absence of IBRD during the glacials could indicate the presence of stabilizing ice shelf systems before retreat in the Wilkes Subglacial Basin, and the relative lack of IBRD during interglacials could reflect the formation of large ice shelves within a deglaciated embayment. However, this scenario would still imply a shift in glacial dynamics and ice sheet extent through a glacial cycle, which is supported by the Nd isotope evidence that pulses in IBRD were associated with major switches in erosional sources (Fig. 2). Further considerations regarding the IBRD proxy at this site have been discussed in detail elsewhere^{6,52}.

Sortable silt measurements. Sortable silt analysis was carried out on 75 samples following an established method⁵³ and reported as the mean grain size of the 10–63 μm detrital fraction (Supplementary Table 4). Carbonate and biogenic silica were removed from sediment samples by reacting with 1 M acetic acid for 24 h, followed by heating to 85 °C in 2 M sodium carbonate for 5 h, with samples being agitated several times during each step. Samples were then suspended in 0.2% sodium hexametaphosphate solution and placed on a rotating wheel before analysis using a Beckman Multisizer 3 Coulter Counter at the University of Queensland. Repeated analysis was performed on a subset of samples in an arbitrary order over several days and average standard deviation of replicate analysis was $\pm 0.1 \mu\text{m}$.

Sedimentological constraints on sediment transport. Our sortable silt record indicates a very fine grain size (13.5 μm), at the low end of the proxy range, and a lack of variation through time (1 s.d. = 0.5 μm) (Supplementary Table 4). While this result is suggestive of a low current speed, we caution that there are caveats in applying the sortable silt method in a setting influenced by turbidity currents and IBRD, and that this proxy cannot provide a quantitative estimate of current speed⁵⁴. Nevertheless, the lack of correlation between sortable silt grain size and detrital Nd isotopes provides no evidence to suggest a control on our provenance record by current transport. In this context, it is important to emphasize that Site U1361 was cored on a turbidite levee deposit²⁰. From both its geometry²⁰ and seafloor bedforms⁵⁵, it is unambiguous that the majority of sediment supplied to the site is derived from the adjacent Jussieu Canyon system, which itself is fed by sediment from the adjacent continental shelf. The lack of variation in the sortable silt grain size record is consistent with such a setting. Given the proximity to such local sediment inputs, and with regional dust inputs at the East Antarctic margin being among the lowest in the world⁵⁶, we also rule out any possible control of dust supply on sediment provenance at Site U1361.

Measurement of Ba/Al ratios. Bulk major element compositions were measured using an Avaatech X-ray fluorescence (XRF) core scanner at the IODP Gulf Coast Repository at Texas A & M University (USA) during March 2011. Non-destructive XRF core scanning was performed at 10 kV in order to measure the relative content of elements including aluminium (Al), iron (Fe), and barium (Ba). Measurements were made continuously at every 5 cm down core from sections U1361A-1H-1 to U1361A-1H-5, over a 1.2-cm² area, with a slit size of 10 mm, a current of 0.8 mA, and a sampling time of 45 s. The Ba and Al counts, and the Ba/Al count ratio, are reported here (Supplementary Table 5).

Interpretation and principal component analysis of Ba/Al ratios. Barium-based proxies (for example, Ba/Al ratios) in pelagic sediments have variously been interpreted as a marine productivity proxy^{57,58}, as a meltwater tracer⁵⁹, or as an indicator of intense bottom currents⁶⁰, among other processes⁶¹. In order to interpret the Ba/Al ratio at Site U1361, we applied principal component analysis (PCA)⁶⁰ to the data from the upper 7 mbsf of the core. The PCA yielded two significant components (Supplementary Table 6). The first principal component (PC1) describes 31.4% of the total variance, with main negative loadings for Ca and Mn, and positive loadings for all other elements (except K). This variable could represent a link between Mn and Ca enrichment peaks during glacial to interglacial transitions, which is the focus of ongoing research. The second principal component (PC2) describes 26.2% of the variance, with positive loadings for Ca, Ba, S, and Si (in descending order), and negative loadings for Fe, Al, K, Mn, and Ti. Such a relationship between the loadings for PC2 is characteristic of biogenic components in the positive axis and aluminosilicates in the negative axis⁶².

Visual inspections suggest that Si and Ca enrichment are related to sediments rich in diatoms (Si) and coccoliths (Ca). Because Ba is related positively to Ca, Si, and S for PC2, we conclude that Ba is related to marine productivity and is probably present in the form of biogenic barite (BaSO₄), while normalization to a detrital immobile element (for example, Al) corrects for dilution effects⁶³. Through time,

loadings of PC2 are virtually parallel to the Ba/Al ratio (Supplementary Table 6), with both showing regular alternations coincident with glacial-interglacial cycles. Based on these correlations, and previous studies at this site^{5,52}, we conclude that the Ba/Al ratio is a robust productivity proxy at U1361.

Diatom counts. Quantitative diatom slides were prepared using a modification of a published method⁶⁴ for 70 samples between 0.05 and 7.45 mbsf at ~10 cm intervals. For each sample, a uniform mass of 0.50 mg raw sediment was digested in 15 ml 30% hydrogen peroxide with 5 ml sodium hexametaphosphate for one hour in a 60 °C water bath, to remove organic material and aid disaggregation. Samples were then triple-rinsed by centrifuging at 1,600 rpm for 7 min, decanting, and topping up with distilled water. After the final rinse, samples were re-suspended, poured into a beaker that was topped up to 100 ml with distilled water, and stirred to homogenize. 300 μl of homogenized solution was pipetted onto a 22 × 38 mm cover slip placed at the bottom of a glass Petri dish, and again topped up with distilled water. Each dish was fitted with a cotton string for wicking and allowed to settle and dry for 48 h, after which the dry cover slips were mounted to glass slides using Norland Optical Adhesive #61 (refractive index of 1.56) and cured under UV light for 30 min.

Three slides were prepared for each sample and examined microscopically. Where possible, 250–300 valves were counted for each slide to yield statistically robust assemblage data. Where total valve counts were lower, at least 35 fields of view were evaluated for each slide to determine absolute diatom abundance (ADA). For each sample, triplicate counts were combined to produce a single data set (Supplementary Table 7). ADA was calculated from a measured field of view of 3.14 mm² and a total settling area of 6,361.72 mm². Diatoms were identified to the species level, following the same taxonomic concepts used during IODP Expedition 318²⁰.

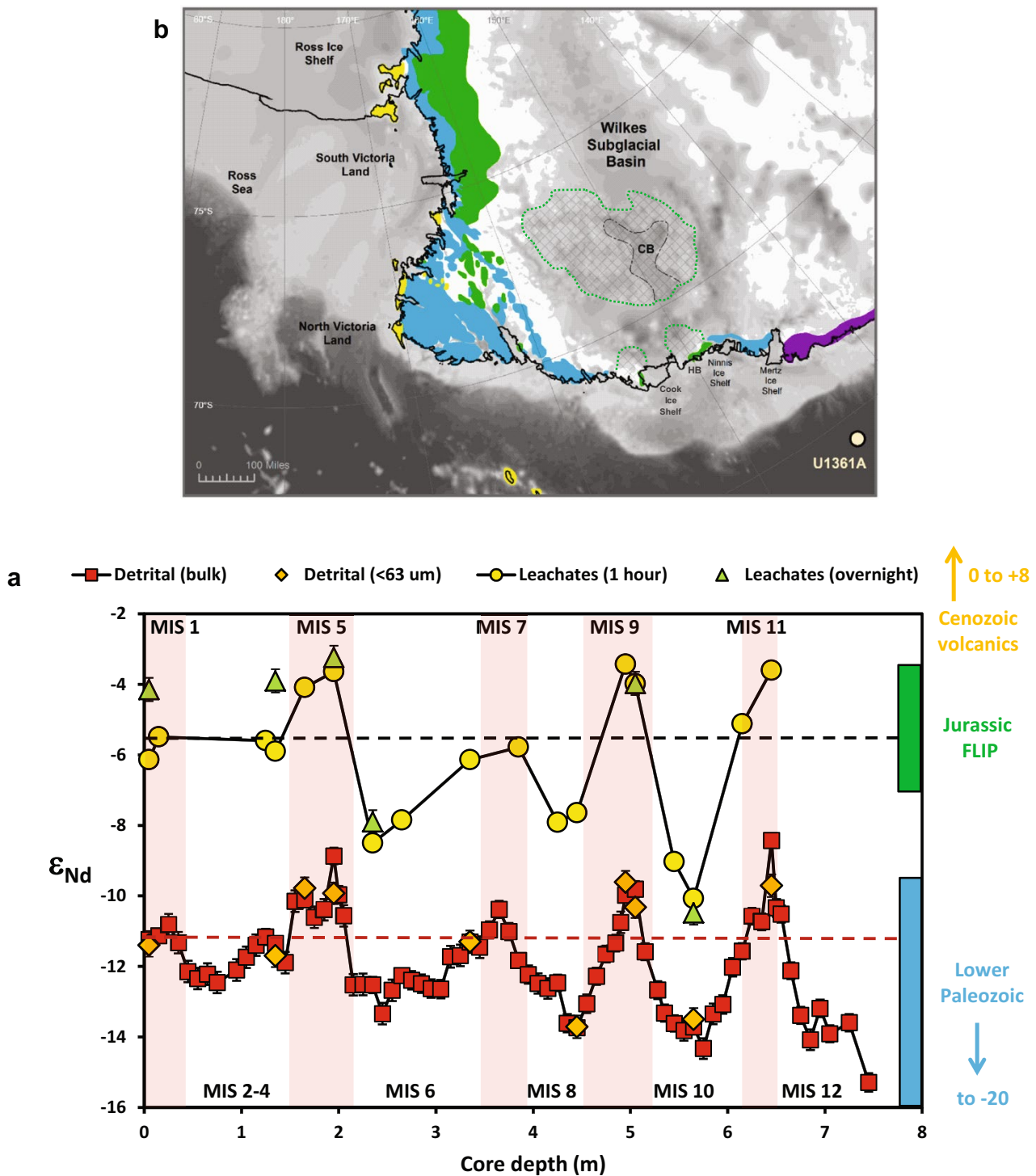
Age control. A Holocene age for the upper portion of U1361A is confirmed by a single radiocarbon date from 0.25–0.27 mbsf (Supplementary Table 8). The radiocarbon analysis was performed on the bulk organic fraction at Beta Analytic. The sample was sieved at 180 μm and pretreated with acid washes to remove carbonate. Conventional ¹⁴C ages were calculated as described⁶⁵ using a $\delta^{13}\text{C}$ correction for isotopic fractionation. A marine reservoir correction of 1300 years was applied⁶⁶ and the ¹⁴C age was calibrated using the Marine 13 calibration curve⁶⁷.

Additional age control for the late Pleistocene interval is provided by the well-established last common occurrence of the diatom *Hemidiscus karstenii* at 190 kyr BP^{68,69}, which is identified at 3.45 mbsf (Supplementary Tables 7 and 8; Extended Data Fig. 3). Immediately below our interval of focus, we identify two palaeomagnetic reversals and two additional diatom last occurrences that further constrain the age of our sequence (Supplementary Table 8; Extended Data Fig. 3). The upper depth constraint for chron C1n (11.19 mbsf) has been revised from published data⁷⁰ based on magnetostratigraphy of the U1361 composite splice (U1361A and U1361B)²⁰. To more precisely constrain the core depths of last occurrences of *Actinocyclus ingens* and *Thalassiosira fasciculata*⁷¹, additional samples from U1361A core 2H were examined following the IODP Expedition 318 shipboard methodology for diatom biostratigraphy²⁰. These secondary constraints confirm that the base of our record is younger than 500 kyr and support our assignment of peaks in multiple proxies to interglacial marine isotope stages 1–11 (Extended Data Fig. 3).

Data availability. All data from this study can be found in the Supplementary Information.

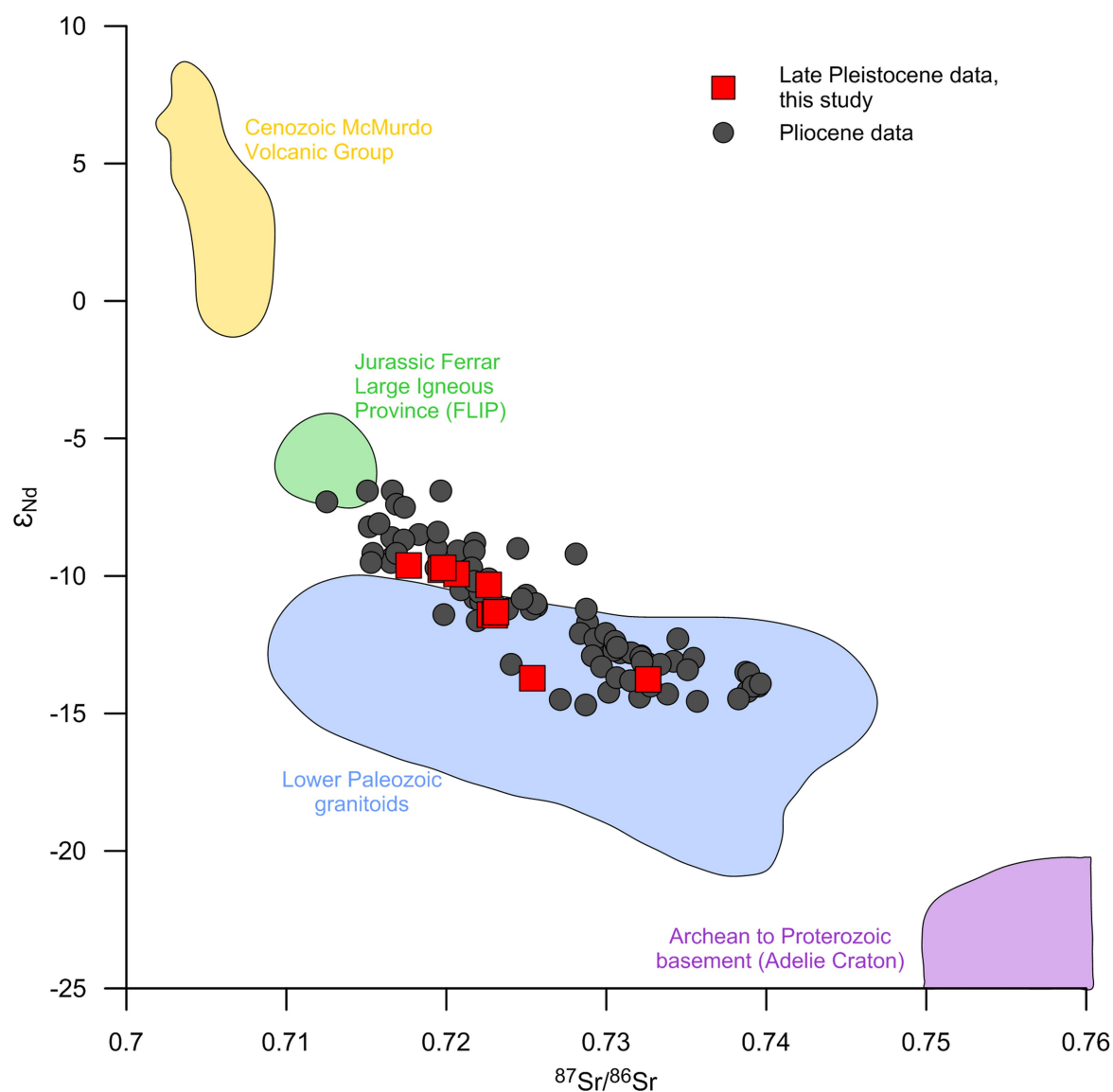
32. Wilson, D. J., Piotrowski, A. M., Galy, A. & Clegg, J. A. Reactivity of neodymium carriers in deep sea sediments: Implications for boundary exchange and paleoceanography. *Geochim. Cosmochim. Acta* **109**, 197–221 (2013).
33. Bayon, G. et al. An improved method for extracting marine sediment fractions and its application to Sr and Nd isotopic analysis. *Chem. Geol.* **187**, 179–199 (2002).
34. Chester, R. & Hughes, M. J. A chemical technique for the separation of ferro-manganese minerals, carbonate minerals and adsorbed trace elements from pelagic sediments. *Chem. Geol.* **2**, 249–262 (1967).
35. Tanaka, T. et al. JNdi-1: a neodymium isotopic reference in consistency with LaJolla neodymium. *Chem. Geol.* **168**, 279–281 (2000).
36. Weis, D. et al. High-precision isotopic characterization of USGS reference materials by TIMS and MC-ICP-MS. *Geochem. Geophys. Geosyst.* **7**, Q08006 (2006).
37. Eisenhauer, A. et al. Grain size separation and sediment mixing in Arctic Ocean sediments: evidence from the strontium isotope systematic. *Chem. Geol.* **158**, 173–188 (1999).
38. Maggi, F. The settling velocity of mineral, biomineral, and biological particles and aggregates in water. *J. Geophys. Res. Oceans* **118**, 2118–2132 (2013).
39. Cook, C. P. et al. Glacial erosion of East Antarctica in the Pliocene: A comparative study of multiple marine sediment provenance tracers. *Chem. Geol.* **466**, 199–218 (2017).
40. Chen, T. Y., Frank, M., Haley, B. A., Gutjahr, M. & Spielhagen, R. F. Variations of North Atlantic inflow to the central Arctic Ocean over the last 14 million years inferred from hafnium and neodymium isotopes. *Earth Planet. Sci. Lett.* **353–354**, 82–92 (2012).

41. Du, J. H., Haley, B. A. & Mix, A. C. Neodymium isotopes in authigenic phases, bottom waters and detrital sediments in the Gulf of Alaska and their implications for paleo-circulation reconstruction. *Geochim. Cosmochim. Acta* **193**, 14–35 (2016).
42. Bayon, G., German, C. R., Burton, K. W., Nesbitt, R. W. & Rogers, N. Sedimentary Fe–Mn oxyhydroxides as paleoceanographic archives and the role of aeolian flux in regulating oceanic dissolved REE. *Earth Planet. Sci. Lett.* **224**, 477–492 (2004).
43. Blaser, P. et al. Extracting foraminiferal seawater Nd isotope signatures from bulk deep sea sediment by chemical leaching. *Chem. Geol.* **439**, 189–204 (2016).
44. Elmore, A. C., Piotrowski, A. M., Wright, J. D. & Scrivner, A. E. Testing the extraction of past seawater Nd isotopic composition from North Atlantic deep sea sediments and foraminifera. *Geochem. Geophys. Geosyst.* **12**, Q09008 (2011).
45. van de Flierdt, T. et al. Neodymium in the oceans: a global database, a regional comparison and implications for palaeoceanographic research. *Philos. Trans. R. Soc. A* **374**, 20150293 (2016).
46. Skinner, L. C. et al. North Atlantic versus Southern Ocean contributions to a deglacial surge in deep ocean ventilation. *Geology* **41**, 667–670 (2013).
47. Frederick, B. C. et al. Distribution of subglacial sediments across the Wilkes Subglacial Basin, East Antarctica. *J. Geophys. Res. Earth Surf.* **121**, 790–813 (2016).
48. Kriisek, L. A. in *Proc. ODP, Sci. Results, 145*, (eds. Rea, D. K., Basov, I. A., Scholl, D. W. & Allan, J. F.) 179–194 (Ocean Drilling Program, College Station, Texas, 1995).
49. Depoorter, M. A. et al. Calving fluxes and basal melt rates of Antarctic ice shelves. *Nature* **502**, 89–92 (2013).
50. Weber, M. E. et al. Millennial-scale variability in Antarctic ice-sheet discharge during the last deglaciation. *Nature* **510**, 134–138 (2014).
51. Stuart, K. M. & Long, D. G. Tracking large tabular icebergs using the SeaWinds Ku-band microwave scatterometer. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **58**, 1285–1300 (2011).
52. Bertram, R. A. et al. Pliocene deglacial event timelines and the biogeochemical response offshore Wilkes Subglacial Basin, East Antarctica. *Earth Planet. Sci. Lett.* **494**, 109–116 (2018).
53. McCave, I. N. Sedimentary processes and the creation of the stratigraphic record in the Late Quaternary North Atlantic Ocean. *Phil. Trans. R. Soc. Lond. B* **348**, 229–241 (1995).
54. McCave, I. N. & Hall, I. R. Size sorting in marine muds: Processes, pitfalls, and prospects for paleoflow-speed proxies. *Geochem. Geophys. Geosyst.* **7**, Q10N05 (2006).
55. Donda, F., Brancolini, G., De Santis, L. & Trincardi, F. Seismic facies and sedimentary processes on the continental rise off Wilkes Land (East Antarctica): evidence of bottom current activity. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **50**, 1509–1527 (2003).
56. Mahowald, N., Albani, S., Engelstaedter, S., Winckler, G. & Goman, M. Model insight into glacial-interglacial paleodust records. *Quat. Sci. Rev.* **30**, 832–854 (2011).
57. Jimenez-Espejo, F. J. et al. Detrital input, productivity fluctuations, and water mass circulation in the westernmost Mediterranean Sea since the Last Glacial Maximum. *Geochem. Geophys. Geosyst.* **9**, Q11U02 (2008).
58. Bonn, W. J., Ginge, F. X., Grobe, H., Mackensen, A. & Fütterer, D. K. Palaeoproductivity at the Antarctic continental margin: opal and barium records for the last 400 ka. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **139**, 195–211 (1998).
59. Plewa, K., Meggers, H. & Kasten, S. Barium in sediments off northwest Africa: A tracer for paleoproductivity or meltwater events? *Paleoceanography* **21**, PA2015 (2006).
60. Bahr, A. et al. Deciphering bottom current velocity and paleoclimate signals from contourite deposits in the Gulf of Cadiz during the last 140 kyr: An inorganic geochemical approach. *Geochem. Geophys. Geosyst.* **15**, 3145–3160 (2014).
61. Griffith, E. M. & Paytan, A. Barite in the ocean—occurrence, geochemistry and palaeoceanographic applications. *Sedimentology* **59**, 1817–1835 (2012).
62. van den Berg, B. C. J. et al. Astronomical tuning for the upper Messinian Spanish Atlantic margin: Disentangling basin evolution, climate cyclicity and MOW. *Global Planet. Change* **135**, 89–103 (2015).
63. Van der Weijden, C. H. Pitfalls of normalization of marine geochemical data using a common divisor. *Mar. Geol.* **184**, 167–187 (2002).
64. Rathburn, A. E., Pichon, J. J., Ayress, M. A. & DeDecker, P. Microfossil and stable-isotope evidence for changes in Late Holocene palaeoproductivity and palaeoceanographic conditions in the Prydz Bay region of Antarctica. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **131**, 485–510 (1997).
65. Stuiver, M. & Polach, H. A. Discussion: reporting of ^{14}C data. *Radiocarbon* **19**, 355–363 (1977).
66. Ingólfsson, Ö. et al. Antarctic glacial history since the Last Glacial Maximum: an overview of the record on land. *Antarct. Sci.* **10**, 326–344 (1998).
67. Reimer, P. J. et al. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55**, 1869–1887 (2013).
68. Presti, M. et al. Sediment delivery and depositional patterns off Adelie Land (East Antarctica) in relation to late Quaternary climatic cycles. *Mar. Geol.* **284**, 96–113 (2011).
69. Gersonde, R. & Barcena, M. A. Revision of the upper Pliocene - Pleistocene diatom biostratigraphy for the northern belt of the Southern Ocean. *Micropaleontology* **44**, 84–98 (1998).
70. Tauxe, L. et al. Chronostratigraphic framework for the IODP Expedition 318 cores from the Wilkes Land Margin: Constraints for paleoceanographic reconstruction. *Paleoceanography* **27**, PA2214 (2012).
71. Cody, R. et al. Selection and stability of quantitative stratigraphic age models: Plio-Pleistocene glaciomarine sediments in the ANDRILL 1B drillcore, McMurdo Ice Shelf. *Global Planet. Change* **96–97**, 143–156 (2012).
72. Lisiecki, L. E. & Raymo, M. E. A Pliocene-Pleistocene stack of 57 globally distributed benthic $\delta^{18}\text{O}$ records. *Paleoceanography* **20**, PA1003 (2005).



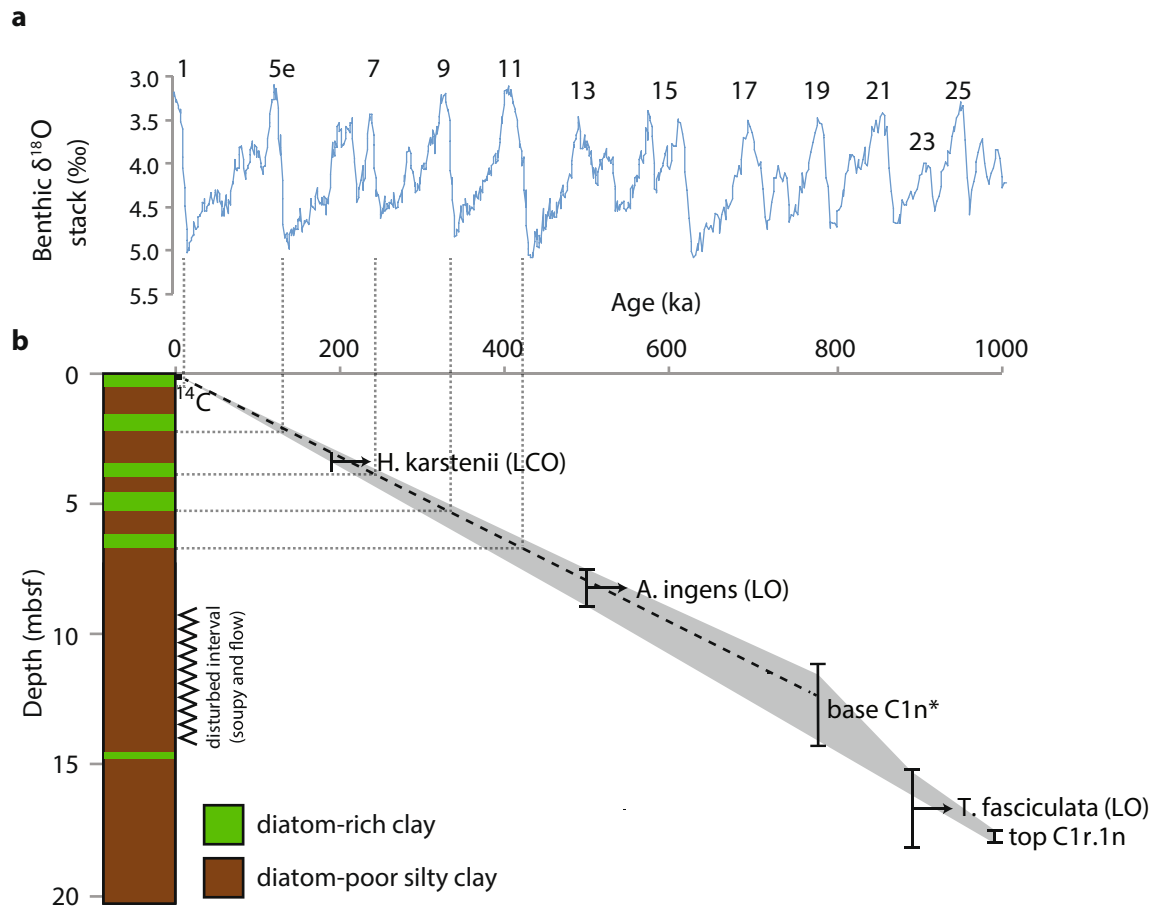
Extended Data Fig. 1 | Neodymium isotope data for bulk detrital sediment, <63 μm fraction, and reductive sediment leachates in U1361A, in comparison to regional bedrock endmembers. a, Down core measurements on the different fractions, with boxes and arrows on the right indicating bedrock endmember compositions in the region (refs ^{5,39} and references cited therein). Horizontal lines indicate Holocene core top values for bulk detrital samples (red dashed line) and 1 h leachate samples (black dashed line). Error bars are 2 s.d. external reproducibility, and are smaller than the symbol sizes where not shown. **b,** Regional bedrock map,

with those same bedrock endmembers located by coloured shading (map redrawn from ref. ⁵, with topography from ref. ¹², and the subglacial extent of the FLIP shown by a green dotted outline inferred from ref. ²¹). In addition to the three endmembers shown in a, purple shading on the map indicates Archaean to Proterozoic basement rocks of the Adélie Craton, with highly unradiogenic Nd isotopic compositions ($\epsilon_{\text{Nd}} = -20$ to -29). CB, Central Basin; HB, Horn Bluff. For interpretation of the leachate and detrital Nd isotope data, see Methods. Map redrawn from ref. ⁵ with permission.



Extended Data Fig. 2 | Neodymium isotope versus Sr isotope crossplot for late Pleistocene fine fraction ($<63\ \mu m$) sediments in U1361A, in comparison to Pliocene detrital sediments from Site U1361 and regional bedrock endmembers. The Pliocene data are based on either

the $<63\ \mu m$ or $<150\ \mu m$ size fractions^{5,39,52}, while bedrock endmember compositions are based on refs^{5,39} (and references cited therein). These data indicate identical trends between the Pliocene and Pleistocene, from which we infer similar provenance variations during both these intervals.



Extended Data Fig. 3 | Age model for U1361A. **a**, LR04 benthic oxygen isotope ($\delta^{18}\text{O}$) stack⁷², labelled with interglacial MIS numbers.

b, Age–depth constraints for U1361A cores 1H and 2H, plotted alongside lithology. Vertical bars for each datum indicate upper and lower depth constraints in U1361A (Supplementary Table 8). Black dashed line is a linear model fit through the Holocene radiocarbon age, *H. karstenii* last common occurrence (LCO), *A. ingens* last occurrence (LO) (upper and lower depths), and the base of chron C1n* (upper depth only, based on the splice to U1361B) (Supplementary Table 8). Forced to an intercept of

0 ka at 0 mbsf, this trendline produces the age–depth equation $y = 64.314x$, where y is age (ka; kyr ago) and x is depth (mbsf). This equation was used to calculate ages for Fig. 3d, e. Grey dotted lines tie lithological transitions to MIS boundaries, based on our age–depth constraints. Note that the Pleistocene section of the core below MIS 12 is affected by sediment disturbance, with extreme disturbance from 9.0–11.67 mbsf (soupy) and 11.67–14.26 mbsf (flow) represented schematically with a zigzag line. We have therefore restricted our provenance study to the upper ~7.5 mbsf.

Ancient herders enriched and restructured African grasslands

Fiona Marshall^{1*}, Rachel E. B. Reid¹, Steven Goldstein², Michael Storozum³, Andrew Wreschnig⁴, Lorraine Hu¹, Purity Kiura⁵, Ruth Shahack-Gross⁶ & Stanley H. Ambrose^{7,*}

Grasslands are one of the world's most extensive terrestrial biomes and are central to the survival of herders, their livestock and diverse communities of large wild mammals^{1–3}. In Africa, tropical soils are predominantly nutrient-limited^{4–6} but productive grassy patches in wooded savannah ecosystems^{2,4} grow on fertile soils created by geologic and edaphic factors, megafauna, fire and termites^{4–6}. Mobile pastoralists also create soil-fertility hotspots by penning their herds at night, which concentrates excrement—and thus nutrients—from grazing of the surrounding savannahs^{7–11}. Historical anthropogenic hotspots produce high-quality forage, attract wildlife and increase spatial heterogeneity in African savannahs^{4,12–15}. Archaeological research suggests this effect extends back at least 1,000 years^{16–19} but little is known about nutrient persistence at millennial scales. Here we use chemical, isotopic and sedimentary analyses to show high nutrient and ¹⁵N enrichment in on-site degraded dung deposits relative to off-site soils at five Pastoral Neolithic²⁰ sites (radiocarbon dated to between 3,700 and 1,550 calibrated years before present (cal. BP)). This study demonstrates the longevity of nutrient hotspots and the long-term legacy of ancient herders, whose settlements enriched and diversified African savannah landscapes over three millennia.

Grassy glades—anthropogenic soil nutrient hotspots—on recent herder settlements increase biodiversity at a landscape scale and influence savannah ecosystem structure and function^{4,12–15}. Although the processes creating these glades are well-understood^{7–9,12–15}, the full time-depth of their creation and effects on African savannahs are as yet unexplored. To investigate the longevity of anthropogenic soil nutrient hotspots, we excavated three Pastoral Neolithic sites located west of the Rift Valley in Ntuka, Narok County, Kenya: Indapi Dapo (site code GvJh121), Oloika 1 (GvJh85), Oloika 2 (GvJh86) and sampled two (GvJm44 and GvJm48) at Lukenya Hill, located to the east of the Rift Valley (Fig. 1, Extended Data Fig. 1, Extended Data Table 1). The sites are located in the Loita–Mara–Serengeti ecosystem and Athi–Kapiti plains. Accelerated mass spectrometry radiocarbon dates for the Ntuka sites range from 2,450 to 2,000 cal. BP and the radiocarbon dates from the Lukenya sites range from 3,700 to 1,550 cal. BP, spanning the earliest to the latest phases of the Pastoral Neolithic²⁰ (Extended Data Table 2). Lithic and ceramic technologies²⁰ indicate that the Oloika sites are members of the Elmenteitan tradition of the Pastoral Neolithic; Indapi Dapo and Lukenya sites belong to the Savannah Pastoral Neolithic tradition (Supplementary Information). The archaeological sites are 60–140 m in diameter (Fig. 1f, g, Extended Data Table 1). All of the sites in the Ntuka study area are located in structurally open grassy patches within wooded savannah grassland. Glades, Pastoral Neolithic sites and abandoned modern settlements are visible as well-defined hectare-scale treeless grassy features on the ground and in satellite imagery (Fig. 1e–g, Extended Data Fig. 2).

A visually distinct grey fine-grained sediment layer, 15–30-cm thick (Fig. 1b–d, Extended Data Fig. 1), occurs in four sites (Oloika

1 and 2, Indapi Dapo and GvJm48) and is discontinuous at the oldest site (GvJm44) (Supplementary Note). Micromorphology shows this grey sediment originates from degraded dung (Extended Data Fig. 2). Colour, texture and structure differ between on-site and off-site sediments (Fig. 1b–d). Phytoliths and dung spherulites⁹ are present in Ntuka on-site sediments and absent in off-site sediments (Extended Data Fig. 3).

Particle size analysis demonstrates that on-site sediments are dominated by silt, relative to coarser sandy off-site samples, and that organic matter and carbonate percentages are higher on-site (Extended Data Fig. 4, Extended Data Table 1). Fourier transform infrared spectroscopy shows that opal and calcite are present in on-site samples at Ntuka sites. Opal originates from silt-sized grass phytoliths. Calcite appears in thin sections as dung spherulites or as microspar (Extended Data Fig. 3). Lukenya Hill sediments do not show substantial mineralogical differences from off-site samples. However, inductively coupled plasma mass spectrometry analyses demonstrate that phosphorous, nitrogen, magnesium and calcium are enriched by an order of magnitude in on- versus off-site samples (Fig. 2, Supplementary Table 1). In some cases, calcium concentrations were elevated by 200–1,000% in on-site sediments. These findings are consistent with the enrichment observed in contemporary pastoral settlements (Supplementary Table 2). Nitrogen and carbon isotope ratios are consistently higher in degraded dung deposits than in natural off-site soils, except at Lukenya site GvJm48 (Fig. 3, Supplementary Table 1). Supplementary Table 2 summarizes metadata from Africa regarding nutrient elevation, and the distinctive vegetation and ecology of historical and Iron Age herder corrals.

Our analyses of micromorphology, mineralogy, and chemical and isotopic composition reveal that elevated levels of nutrients persist for 3,000 years in decomposed dung at Neolithic herder sites in the grasslands of southern Kenya. Our interpretations of the archaeological data are based on ethno-archaeological and ecological studies of contemporary pastoral settlements that show enrichment in weight percentage (wt%) N and ¹⁵N of soil organic matter, grass phytoliths, dung spherulites and mineral nutrients (especially phosphorous and calcium), relative to off-site samples^{8,18,21,22}. Nitrogen in cattle and sheep and goat dung is enriched in ¹⁵N because dung is composed of a mixture of both excreted undigested plant material and ¹⁵N-enriched proteinaceous material from the animals themselves^{18,22}. Volatilization of ¹⁵N-depleted ammonia from dung and urine decomposition in semi-arid environments also increases soil ^{δ¹⁵N}^{21,23}. Soil organic ^{δ¹³C} values on archaeological sites are significantly higher than off-site values, and the highest ^{δ¹³C} values are associated with the highest ^{δ¹⁵N} values. This pattern is consistent with soil organic carbon and nitrogen being derived from dung and urine excreted by herbivores that graze on C₄ plants. Phosphorous originates from the organic component of dung^{24,25}. Calcium carbonate is concentrated in decomposed dung in the form of dung spherulites, which elevate calcium levels. Bones are additional sources of phosphate, magnesium and calcium. The presence

¹Department of Anthropology, Washington University in St Louis, St Louis, MO, USA. ²Department of Archaeology, Max Planck Institute for the Science of Human History, Jena, Germany. ³Earth Observatory of Singapore, N2-01a-15, Nanyang Technological University, Singapore, Singapore. ⁴AECOM, Charlotte, NC, USA. ⁵Directorate of Museums, Sites and Monuments, National Museums of Kenya, Nairobi, Kenya. ⁶Department of Maritime Civilizations, Leon Recanati Institute of Maritime Studies, The Leon H. Charney School of Marine Sciences, University of Haifa, Haifa, Israel.

⁷Department of Anthropology, University of Illinois, Urbana, IL, USA. *e-mail: fmarshall@wustl.edu; ambrose@illinois.edu

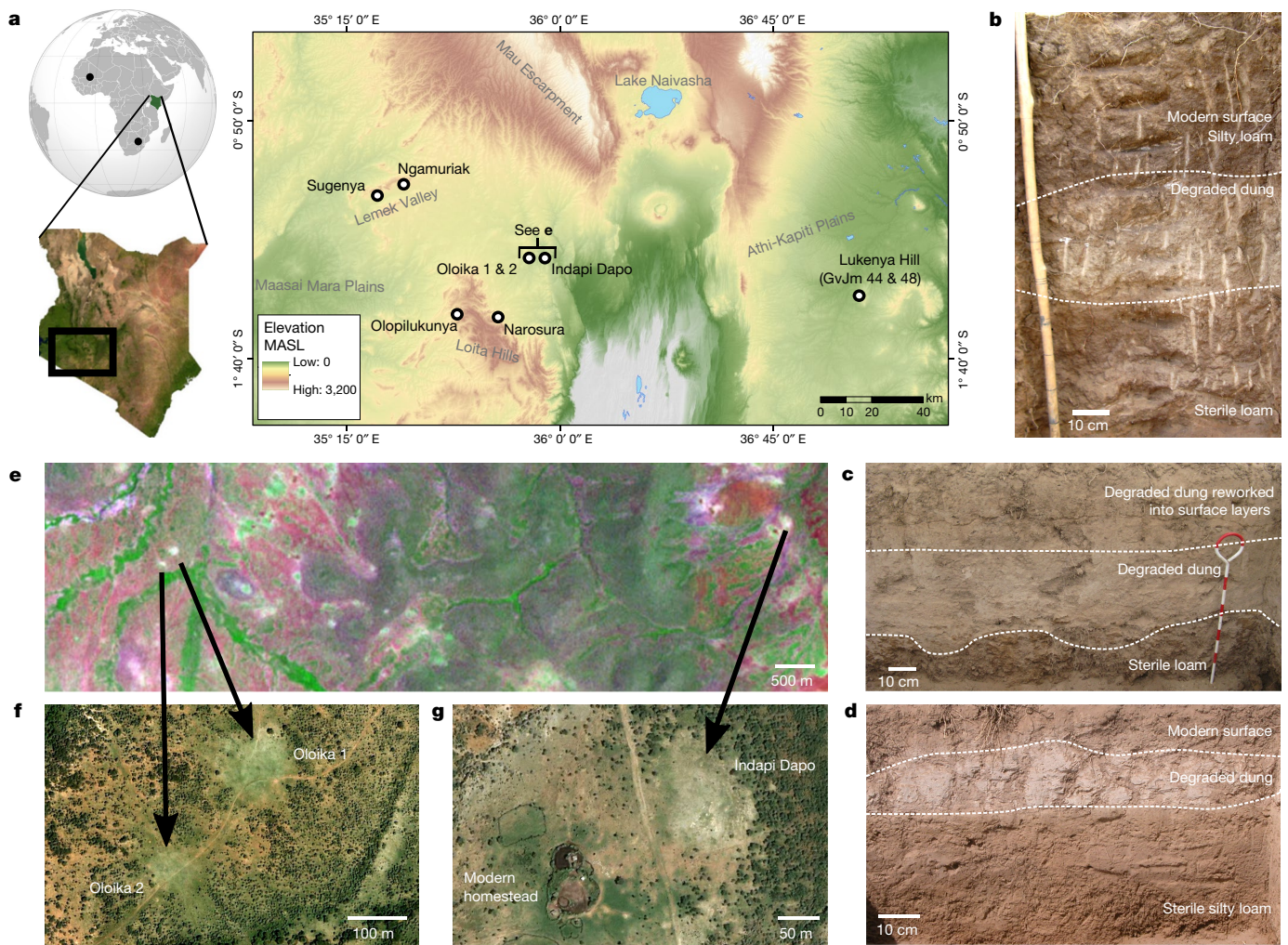


Fig. 1 | Study areas and sampled sites. **a**, Distribution of Pastoral Neolithic sites in south-western Kenya. MASL, metres above sealevel. Digital elevation data from NASA SRTM. **b–d**, Dung deposits visible in profiles at 1,500–3,000-year-old herder settlements at Lukenya Hill (GvJm48, **b**), Oloika 2 (**c**) and Indapi Dapo (**d**). **e**, False-colour LANDSAT-7 image of Narok county Ntuka study area on 3 February 2003,

Oloika 1, Oloika 2 and Indapi Dapo site glades are visible as white patches, and modern Maasai settlements as dark red patches. **f**, **g**, High nutrient legacies encourage open grassy plant succession relative to bushy off-site vegetation at Oloika 1 and Oloika 2 (**f**), and Indapi Dapo (**g**). Globe in **a**, CC BY 2.0 licence. **b–d**, Photographs by F.M. **f**, **g**, Imagery from Google Earth Pro, Digital Globe.

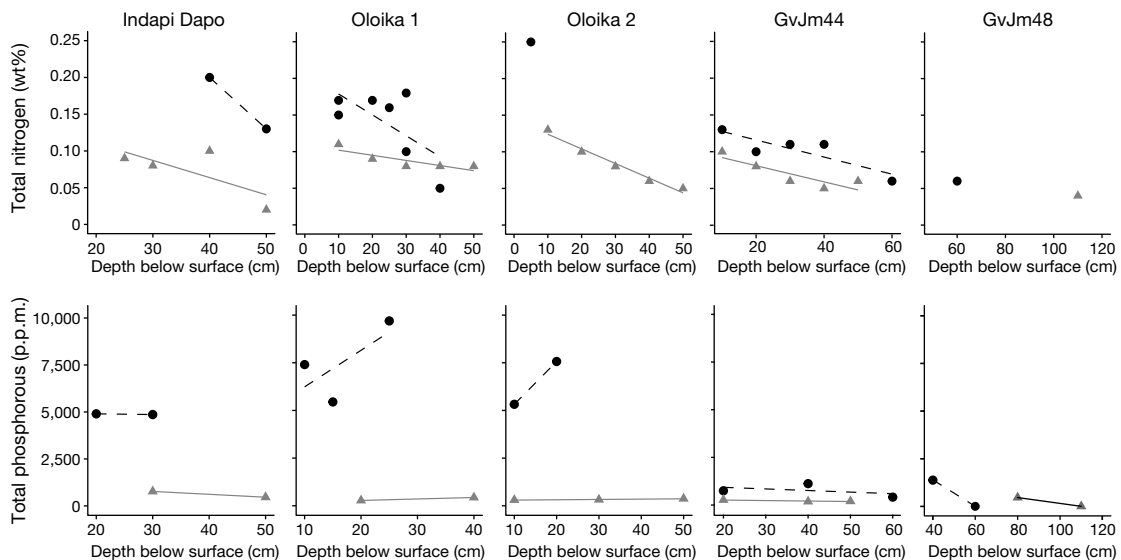


Fig. 2 | Elemental nitrogen and phosphorous concentrations in sediment from on- and off-site stratigraphic sections. Elemental nitrogen (top) and phosphorous (bottom) concentrations in sediments from on-site (black circles, $n = 16$ for N and $n = 12$ for P) and off-site

(grey triangles, $n = 20$ for N and $n = 12$ for P) stratigraphic sections at the sampled archaeological sites. All samples are independent. Least-squares regressions plotted as dashed lines for on-site samples and solid lines for off-site samples.

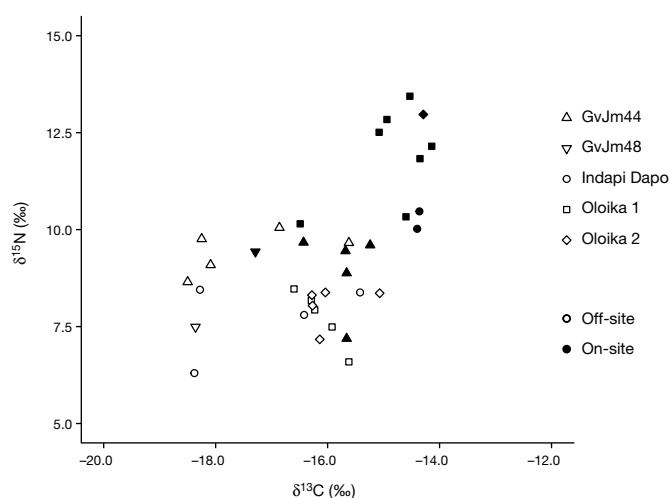


Fig. 3 | $\delta^{13}\text{C}$ and $\delta^{15}\text{N}$ values measured in on-site and off-site sediment samples from five archaeological sites in East Africa. On-site samples, closed symbols ($n = 16$); off-site samples, open symbols ($n = 20$). All samples are independent.

of microspar suggests that dung spherulites dissolved and calcite re-precipitated as microspar. This is supported by the presence of manganese-oxide florets in thin sections, suggesting occasional hydromorphic conditions at some sites. The precipitation of microbially mediated carbonates²² and translocation of carbonate-rich solutions down profiles⁹ resulted in the formation of basal calcitic crusts.

A previous study of a 40-year time series of abandoned Maasai pastoral settlements in Kenya demonstrated that organic matter content declines substantially after about 20–30 years⁹. Mineralogical cascades triggered by the products of organic matter (for example, formation of phosphate minerals) stabilize at this point in diagenesis, sites become more deeply buried and minerals and elements may persist for millennia^{9,25}. Continued enrichment of pastoral corral sediments by wild ungulates and domestic herds attracted to palatable forage, and soil–plant–herbivore feedbacks may contribute to persistence of anthropogenic hotspots^{7,11,15}. Our results, as well as geochemical analyses of the Pastoral Neolithic site of Suganya¹⁸, reveal the persistence of nutrient-enriched dung-derived deposits over three millennia.

These findings reinforce the environmental importance of the fertile grassy patches created by the earliest southern Kenyan pastoralists. Widespread settlements generated nutrient-enriched, hectare-scale microhabitats. Neolithic, Iron Age and recent herder sites are visible in satellite images of the Ntuka area as glades (4,400–15,000 m² in size) (Fig. 1e–g, Extended Data Fig. 2d). Research on the Laikipia Plateau of Central Kenya complements Pastoral Neolithic findings, extending our understanding of nutrient stabilization and glade landscapes into the Pastoral Iron Age. The seventeenth–eighteenth-century-AD settlement of Maili Sita preserves phytoliths, spherulites and elevated nutrients in dung deposits that support characteristic grass species¹⁹. The fifteenth-century-AD Maasai Plains and unexcavated sites on the Laikipia Plateau reveal a broad distribution of 15–45-ha pastoral glades²⁶. Pastoral Neolithic and Iron Age sites in diverse Kenyan savannas demonstrate the spatial influences of niche construction by pastoralists on soil nutrients and savannah heterogeneity, on timescales that range from five centuries to three millennia.

Influences of the settlements of ancient and recent mobile herders in eastern Africa create landscape palimpsests. Recent herders make similar choices about where to locate their settlements to ancient ones with regard to slope and distance from water^{27,28}. As a result, contemporary Maasai herders settle on or near Pastoral Neolithic settlements in the Lemek Valley²⁹ and Ntuka areas of southwest Kenya (Fig. 1g, Extended Data Fig. 5). Such settlement clusters are also attractive to modern herders because of the proximity of nutrient-rich grazing that supports growth and lactation (for example, for calves)^{3,7}. Studies of

Maasai settlements constructed over the last 60–100 years suggest that 1–20% of savannas, and most land near water, has been settled over the past 100 years^{12,27}.

East African findings draw attention to the temporal and spatial scale of pastoral legacies. Metadata on modern and ancient African pastoral settlements indicate that influences on nutrient enrichment and ecology are broadly dispersed (Supplementary Table 2). Ecological research on a South African Iron Age site in the Nylsvley, Nature Reserve Limpopo Province (about 700 years old), compares modern glade formation processes in South Africa, the Sahel and eastern Africa (marked on Fig. 1a) and documents increased biodiversity and forage quality on this ancient anthropogenic hotspot⁴. In the Limpopo Valley and eastern Botswana, corrals in Iron Age settlements (about 1,200–500 years old) (Supplementary Table 2) are characterized by grassy vegetation and eutrophic dung-derived soils^{16,17}. Landscape-scale on- and off-site studies have the potential to resolve ancient patterns at finer scales. However, biogeochemical data on Neolithic and other East African sites, and ecological and nutrient data from widespread Iron Age and historic sites, indicate that herders have had a role in structuring and diversifying African savannah ecosystems for up to three millennia.

Outside Africa, reinforced nutrient enrichment related to Iron Age pastoral activity in arid environments has also been documented in the southern Levant³⁰. Exploration of nutrient enrichment by ancient pastoral settlements in temperate and high-altitude grasslands in Central Asia³¹ and South America³² will yield insights into local and regional variability and the global importance of prehistoric herder influences on nutrient flows and grassland ecology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0456-9>.

Received: 14 September 2017; Accepted: 30 July 2018;

Published online 29 August 2018.

- Reynolds, S. G. in *Grasslands of the World* (eds Suttle, J. M. et al.) 1–17 (Food and Agriculture Organization of the United Nations, Rome, 2005).
- Sankaran, M. et al. Determinants of woody cover in African savannas. *Nature* **438**, 846–849 (2005).
- Reid, R. S. *Savannas of our Birth* (Univ. California Press, Berkeley, 2012).
- Scholes, R. J. & Walker, B. H. *An African Savanna. Synthesis of the Nylsvley study* (Cambridge Univ. Press, Cambridge, 1993).
- Bell, R. H. V. in *Ecology of Tropical Savannas* (eds Huntley, B. J. & Walker, B. H.) 193–216 (Springer, Berlin, 1982).
- Cech, P., Kuster, T., Edwards, P. J. & Venterink, H. O. Effects of herbivory, fire and N₂-fixation on nutrient limitation in a humid African savanna. *Ecosystems* **11**, 991–1004 (2008).
- Augustine, D. J., McNaughton, S. J. & Frank, D. A. Feedbacks between soil nutrients and large herbivores in a managed savanna ecosystem. *Ecol. Appl.* **13**, 1325–1337 (2003).
- Porensky, L. M. & Veblen, K. E. Generation of ecosystem hotspots using short-term cattle corrals in an African savanna. *Rangeland Ecol. Manag.* **68**, 131–141 (2015).
- Shahack-Gross, R., Marshall, F. & Weiner, S. Geo-ethnoarchaeology of pastoral sites: the identification of livestock enclosures in abandoned Maasai settlements. *J. Archaeol. Sci.* **30**, 439–459 (2003).
- Turner, M. Long term effects of daily grazing orbits on nutrient availability in Sahelian West Africa: I. Gradients in the chemical composition of rangeland soils and vegetation. *J. Biogeogr.* **25**, 669–682 (1998).
- Georgiadis, N. J. & McNaughton, S. J. Elemental and fibre contents of savanna grasses: variation with grazing, soil type, season and species. *J. Appl. Ecol.* **27**, 623–634 (1990).
- Muchiru, A. N., Western, D. J. & Reid, R. S. The impact of abandoned pastoral settlements on plant and nutrient succession in an African savanna ecosystem. *J. Arid Environ.* **73**, 322–331 (2009).
- van der Waal, C. et al. Large herbivores may alter vegetation structure of semi-arid savannas through soil nutrient mediation. *Oecologia* **165**, 1095–1107 (2011).
- Muchiru, A. N., Western, D. J. & Reid, R. S. The role of abandoned pastoral settlements in the dynamics of African large herbivore communities. *J. Arid Environ.* **72**, 940–952 (2008).
- Porensky, L. M. & Veblen, K. E. Grasses and browsers reinforce landscape heterogeneity by excluding trees from ecosystem hotspots. *Oecologia* **168**, 749–759 (2012).
- Denbow, J. R. *Cenchrus ciliaris*: an ecological indicator of Iron Age middens using aerial photography in eastern Botswana. *S. Afr. J. Sci.* **75**, 405–408 (1979).

17. Huffman, T. N., Elberg, M. & Watkeys, M. Vitriified cattle dung in the Iron Age of southern Africa. *J. Archaeol. Sci.* **40**, 3553–3560 (2013).
18. Shahack-Gross, R., Simons, A. & Ambrose, S. H. Identification of pastoral sites using stable nitrogen and carbon isotopes from bulk sediment samples: a case study in modern and archaeological pastoral settlements in Kenya. *J. Archaeol. Sci.* **35**, 983–990 (2008).
19. Boles, O. J. C. & Lane, P. The green, green grass of home: an archaeo-ecological approach to pastoralist settlement in central Kenya. *Azania* **51**, 507–530 (2016).
20. Ambrose, S. H. in *Encyclopedia of Prehistory, Vol. 1, Africa* (eds Peregrine, P. N. & Ember, M.) 97–109 (Kluwer/Plenum, New York, 2001).
21. Frank, D. A., Evans, R. D. & Tracy, B. F. The role of ammonia volatilization in controlling the natural ^{15}N abundance of a grazed grassland. *Biogeochemistry* **68**, 169–178 (2004).
22. Macharia, A. N., Uno, K. T., Cerling, T. E. & Brown, F. H. Isotopically distinct modern carbonates in abandoned livestock corrals in northern Kenya. *J. Archaeol. Sci.* **39**, 2198–2205 (2012).
23. Ambrose, S. H. Effects of diet, climate and physiology on nitrogen isotope abundances in terrestrial foodwebs. *J. Archaeol. Sci.* **18**, 293–317 (1991).
24. Shahack-Gross, R. Herbivorous livestock dung: formation, taphonomy, methods for identification, and archaeological significance. *J. Archaeol. Sci.* **38**, 205–218 (2011).
25. Shahack-Gross, R., Berna, F., Karkanas, P. & Weiner, S. Bat guano and preservation of archaeological remains in cave sites. *J. Archaeol. Sci.* **31**, 1259–1272 (2004).
26. Lane, P. An outline of the later Holocene archaeology and precolonial history of the Ewaso Basin, Kenya. *Smithson. Contrib. Zool.* **632**, 11–30 (2011).
27. Lamprey, R. H. & Reid, R. S. Expansion of human settlement in Kenya's Maasai Mara: what future for pastoralism and wildlife? *J. Biogeogr.* **31**, 997–1032 (2004).
28. Western, D. & Dunn, T. Environmental aspects of settlement site decisions among pastoral Maasai. *Hum. Ecol.* **7**, 75–98 (1979).
29. Robertshaw, P., Pilgram, T., Siiriainen, A. & Marshall, F. in *Early Pastoralists of Southwestern Kenya* (ed. Robertshaw, P.) 36–51 (British Institute in Eastern Africa, Nairobi, 1990).
30. Shahack-Gross, R. & Finkelstein, I. Subsistence practices in an arid environment: a geoarchaeological investigation in an Iron Age site, the Negev Highlands, Israel. *J. Archaeol. Sci.* **35**, 965–982 (2008).
31. Frachetti, M. D. *Pastoralist Landscapes and Social Interaction in Bronze Age Eurasia* (Univ. of California Press, Berkeley, 2008).
32. Bruno, M. C. & Hastorf, C. A. in *An Archaeology of Andean Pastoralism* (eds Capriles, J. M. & Tripcevich, N.) 55–65 (Univ. New Mexico Press, Albuquerque, 2016).

Acknowledgements We thank the Kenya Ministry of Science and Technology for permission to conduct research (MOST 13/001/38C234, NCST RRI/12/BS011/38) and National Museums of Kenya for research affiliation, excavation licence and support. We are grateful to J. K. Mulwa and M. Mulwa, for site access at Lukenya and to A. Kabiru, J. M. Munyiri, N. Ole Simpai, H. Ole Saitabau and J. K. Ole Tumpuya for research assistance. Funding was received from Washington University in St Louis I-CARES and support from the Liu and the Kidder laboratories, the Nano Research Facility at Washington University, NSF Grant No. ECS-0335765 and the University of Illinois Environmental Isotope Paleobiogeochemistry Laboratory.

Reviewer information Nature thanks N. Boivin, R. Conant, J. Lee-Thorp and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions F.M. and S.H.A. designed the study; F.M., S.H.A., A.W. and S.G. collected field data, R.E.B.R., R.S.-G., S.G., M.S., S.A. and L.H. performed analyses. All authors discussed data and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0456-9>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0456-9>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to F.M. or S.H.A.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample sizes.

Soil profiles and samples from excavations. Sites were identified in the Ntuka region during site surveys by S.H.A. before 2011. Surface exposures and animal burrows in Pastoral Neolithic sites were examined in 2011 for potential corral areas. Excavation proceeded by 5-cm spits or natural levels, where detected. All finds were screened through 5-mm mesh. Samples were taken within excavations for micromorphology, flotation and sediment analyses. Off-site samples were taken 30–40 m from the edges of archaeological sites, in bush-dominated microhabitats.

Sediment analyses. Elemental analysis of bulk sediment samples used an Agilent 7750 ICP-MS. Dried samples (~0.1 g) were digested in concentrated HNO₃ in a Mars-6 microwave digester at 180 °C for one hour. After digestion, the supernatant was diluted to a 10× solution, filtered through a 0.22-μm filter and diluted again to make a final 100× solution. Internal reference standards and HNO₃ blanks were used to calibrate the ICP-MS. Potential memory effects were monitored by running blanks after every 5 samples and drift monitored by running internal standards after every 15 samples. Particle size, loss on ignition, magnetic susceptibility and chemical composition analyses were conducted at the Geoarchaeology and Nano facilities at Washington University in St. Louis. Micromorphology and FTIR analyses were performed at the Laboratory for Sedimentary Archaeology, University of Haifa (see Supplementary Information).

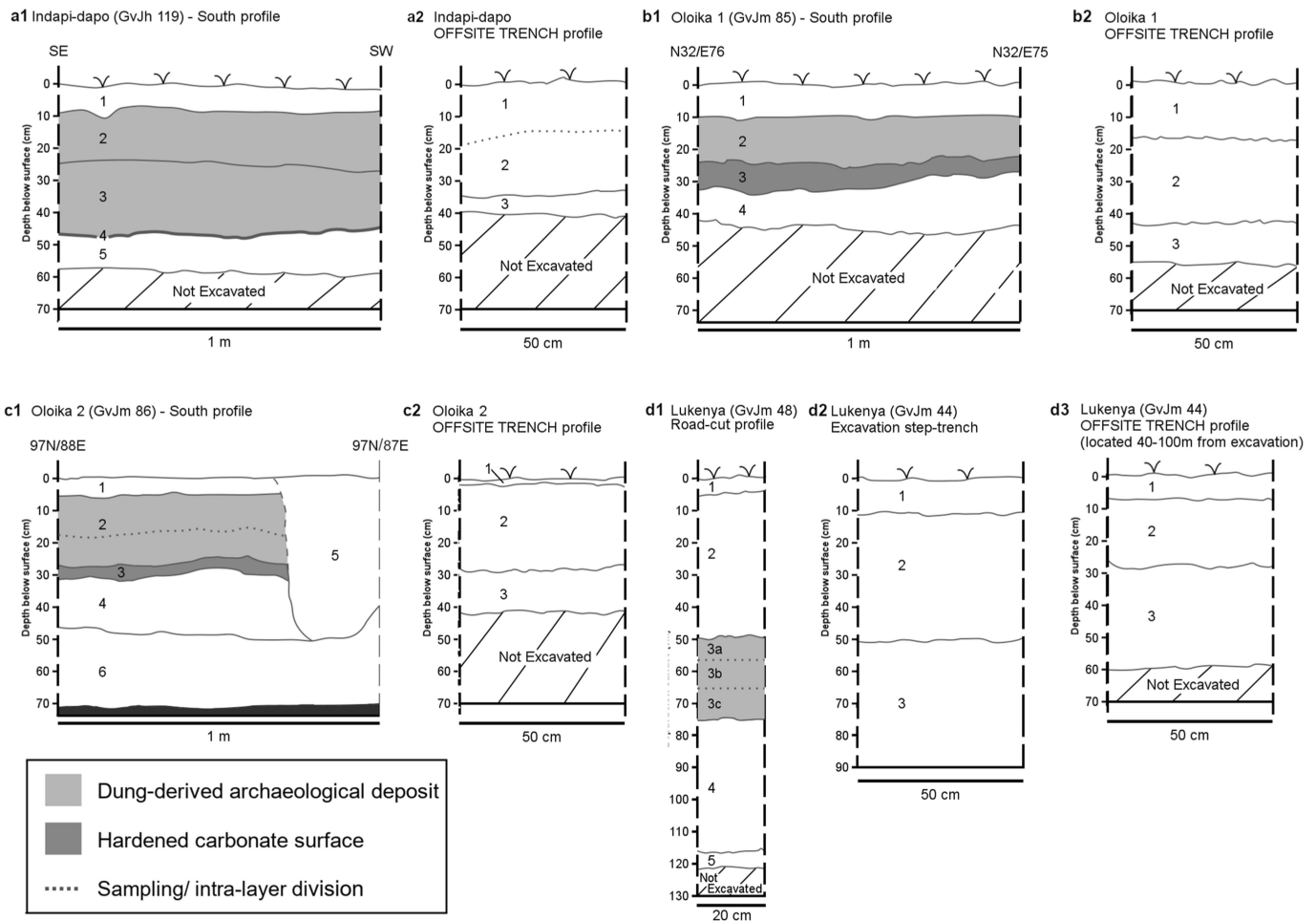
Stable isotopes. Analyses on bulk sediment samples were carried out at Washington University in St. Louis, using a Flash 2000 elemental analyser coupled to a Thermo Delta V Plus continuous-flow isotope ratio mass spectrometer. Samples were homogenized in an agate mortar, and pestle-and-sieved to a particle size of <250 μm. We aimed to analyse at least 20 μg of nitrogen, so 15–80 mg of sediment was weighed into 5 × 9-mm tin capsules. Samples for carbon isotope analysis were treated to remove carbonates with 2 M HCl until effervescence ceased (~24 h), rinsed 5 times with MQ water, dried in a 70 °C oven, and weighed into 5 × 9-mm tin capsules. Our results are expressed as δ¹⁵N and δ¹³C as parts per thousand (‰) relative to AIR and Vienna Pee Dee Belemnite standards, respectively. The average analytical precision for both C and N was <0.2 ‰ based on the standard deviation of 24 replicates of an in-house standard (Bob's Red Mill millet flour) and 18 replicates of a second in-house standard (acetanilide). Weight percentage C and N are estimated based on standards of known elemental composition; wt% precision of these known compounds is better than 0.1%. We used the lme4 package in R to perform a linear mixed effects analysis of the relationship between sediment isotope values and the on-site presence of a dung profile (see Supplementary Information).

Radiocarbon dating. Faunal collagen and enamel apatite samples from Indapi Dapo and Oloika 2 (Extended Data Table 2) were prepared for radiocarbon dating at the Environmental Isotope Paleobiogeochemistry Laboratory, Department of Anthropology, University of Illinois, Urbana, and the Radiocarbon Laboratory of the Illinois State Geological Survey at the University of Illinois. To purify collagen, dentine was demineralized using 0.2 M HCl, rinsed 8× with dH₂O, treated with 0.125 M NaOH to removed soil organic contaminants, rinsed 8× with dH₂O and hydrolysed at 70 °C in 10–3 M HCl. Freeze-dried collagen was converted to CO₂ using sealed-tube combustion, and cryogenically distilled for AMS dating at the UC Irvine Accelerator Mass Spectrometry radiocarbon laboratory. Enamel was separated from dentine and cementum and ground in an agate mortar. A ~400-mg sample was treated with 25 ml 2.63% NaOCl (sodium hypochlorite) to remove organic matter, rinsed 8× with dH₂O, and reacted with 25 ml 0.1 M acetic acid under vacuum to remove adsorbed and diagenetic carbonate, alternating with return to atmospheric pressure with CO₂-free N₂. Cycling between vacuum and N₂ continued at ~15–30-min intervals until the bubbling reaction ceased (~3–4 h). Samples were rinsed 5× in distilled water and freeze dried. Purified samples were reacted with 100% H₃PO₄ and CO₂ from structural carbonate was purified by cryogenic distillation for AMS dating.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. All data generated or analysed in this study are included in the paper and its Supplementary Information. Site descriptions are in Supplementary Information, radiocarbon dates in Extended Data Table 2 and particle size, ICPMS, FTIR, isotope and micromorphology data in Supplementary Table 1. Remaining soil samples are curated in the Archaeology Division of the National Museums of Kenya, Nairobi.

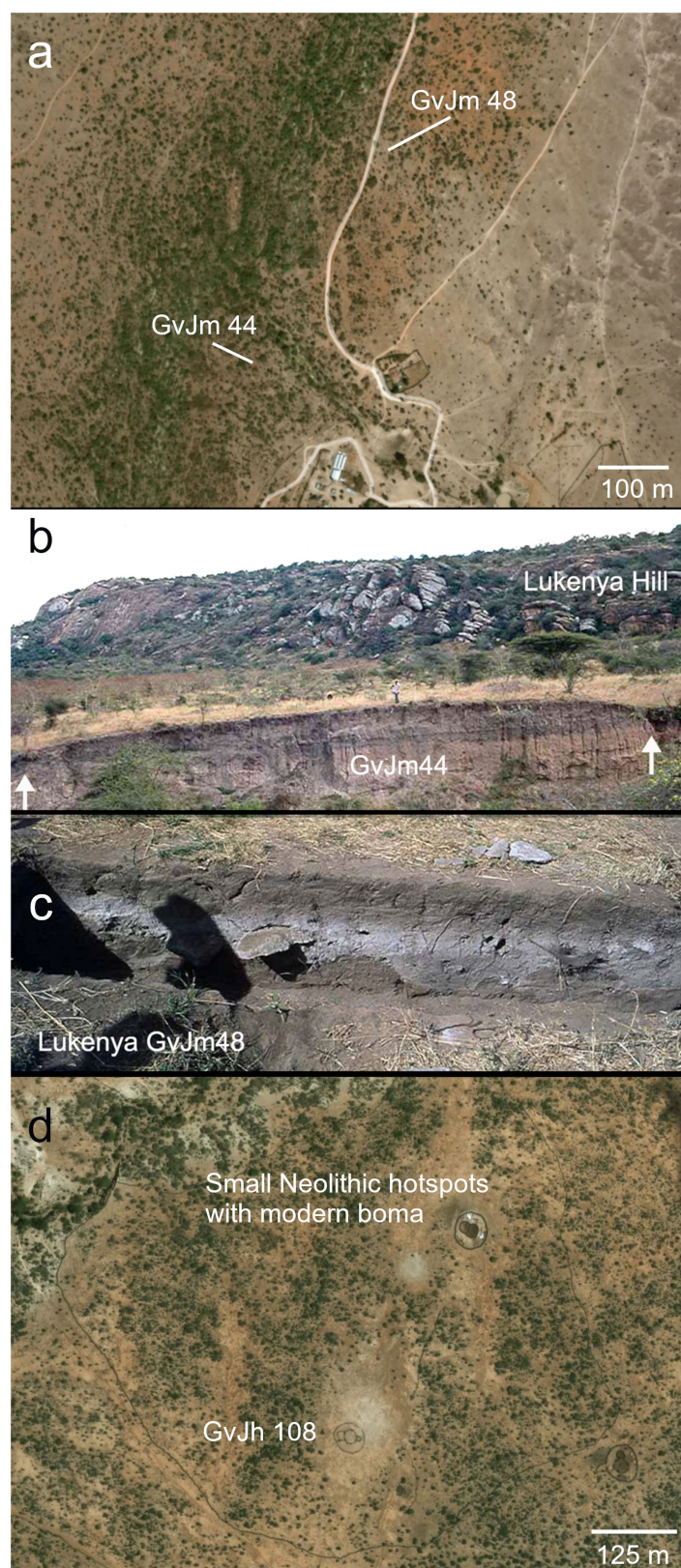
33. Wickham, H. *ggplot2: Elegant Graphics for Data* (Springer, New York, 2009).
34. Bronk Ramsey, C. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**, 227–260 (2009).
35. Hogg, A. G. et al. SHCal13 Southern Hemisphere calibration, 0–50,000 years cal. bp. *Radiocarbon* **55**, 1889–1903 (2013).
36. Simons, A. *The Development of Early Pastoral Societies in South-Western Kenya: A Study of the Faunal Assemblage from Suganya and Oldorotua 1*. PhD thesis, La Trobe Univ. (2004).
37. Nelson, C. M. & Kimengich, J. in *Origin and Early Development of Food-Producing Cultures in Northeast Africa* (ed. Krzyzaniak, L.) 481–487 (Polish Academy of Sciences, Poznan, 1984).
38. Bower, J. R. F., Nelson, C. M., Waibel, A. F. & Wandibba, S. The University of Massachusetts' Later Stone Age/Pastoral Neolithic comparative study in central Kenya: an overview. *Azania* **12**, 119–146 (1977).



Extended Data Fig. 1 | On-site versus off-site sediment profiles for sampled locales. **a**, On-site and off-site stratigraphy at Indapi Dapo.

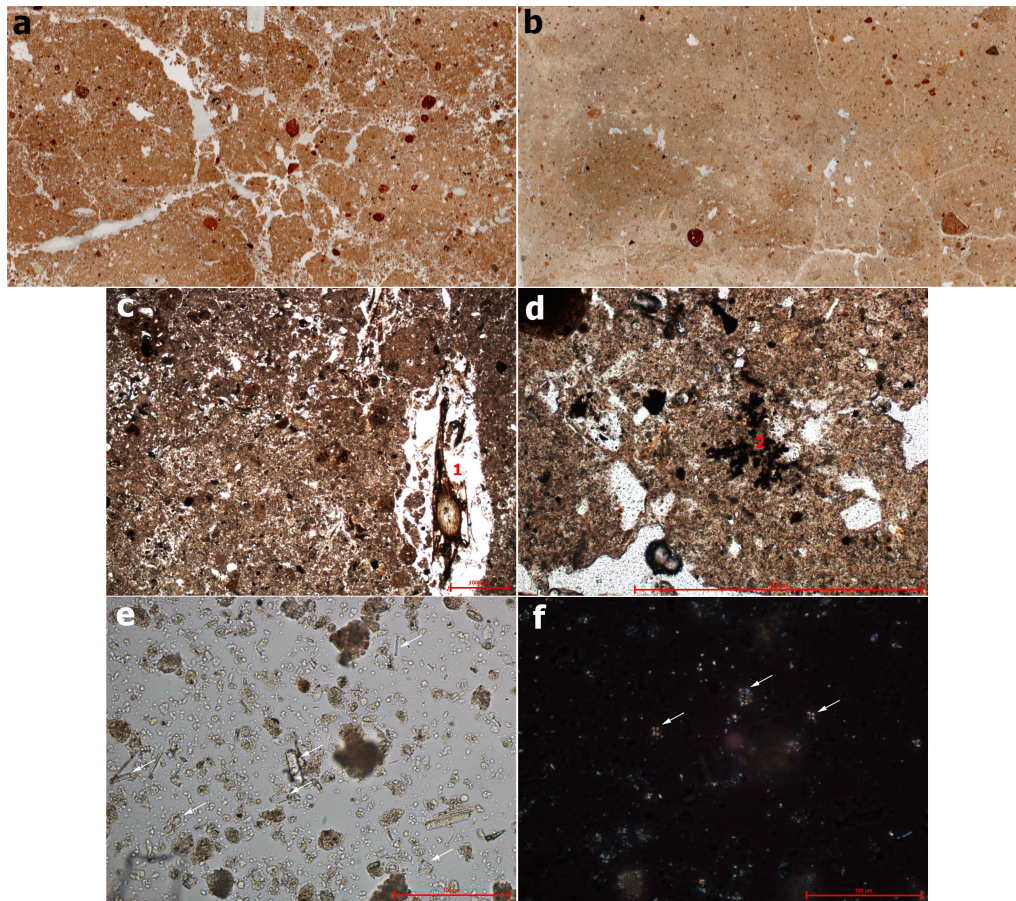
a1 depicts on-site stratigraphy: (1) modern topsoil, (2) light grey cultural horizon, (3) light yellow-brown cultural and dung horizon, (4) discontinuous harder trampled surface and (5) dark yellow-brown sterile sediments. **a2** depicts off-site stratigraphy: (1) loamy modern topsoil, (2) brown silts with carbonate nodules and (3) rocky bedrock-derived sediment. **b**, On-site and off-site stratigraphy at Oloika 1. **b1** depicts on-site stratigraphy: (1) modern topsoil, (2) pale grey cultural and dung horizon, (3) compacted cultural horizon with hard undulating calcium carbonate crust and (4) sterile oxidized palaeosol with manganese nodules. **b2** depicts off-site stratigraphy: (1) light brown modern topsoil, (2) grey-brown sediment with carbonate nodules, (3) oxidized subsoil. **c**, On-site and off-site stratigraphy at Oloika 2. **c1** depicts on-site stratigraphy: (1) modern topsoil, (2) pale grey cultural and dung horizon,

(3) compacted calcium carbonate lens, (4) oxidized subsoil, (5) recent animal burrow and (6) oxidized subsoil pisolithic formation with manganese nodules. **c2** depicts off-site stratigraphy: (1) light brown modern topsoil, (2) grey-brown sediment with carbonate nodules and (3) consolidated lighter grey soil with increasing carbonate nodules. **d**, On-site road-cut (GvJm48) and step-trench (GvJm44) stratigraphy, and off-site stratigraphy at GvJm44. **d1** depicts the GvJm48 road-cut stratigraphy: (1) modern topsoil, (2) grey-brown silty loam, (3a, 3b, 3c) top, middle and bottom, respectively, of pale grey silty loam cultural and dung horizon, (4) pre-cultural loam palaeosol and (5) bedrock-derived weathered sediments. **d2** depicts the GvJm44 step-trench stratigraphy: (1) modern topsoil, (2) dark yellow-brown clay grading to silty loam cultural horizon and (3) lower dark brown silty loam cultural horizon. **d3** depicts off-site stratigraphy at GvJm44: (1) modern topsoil, (2) dark brown to red brown sandy loam (3) sandy loam.



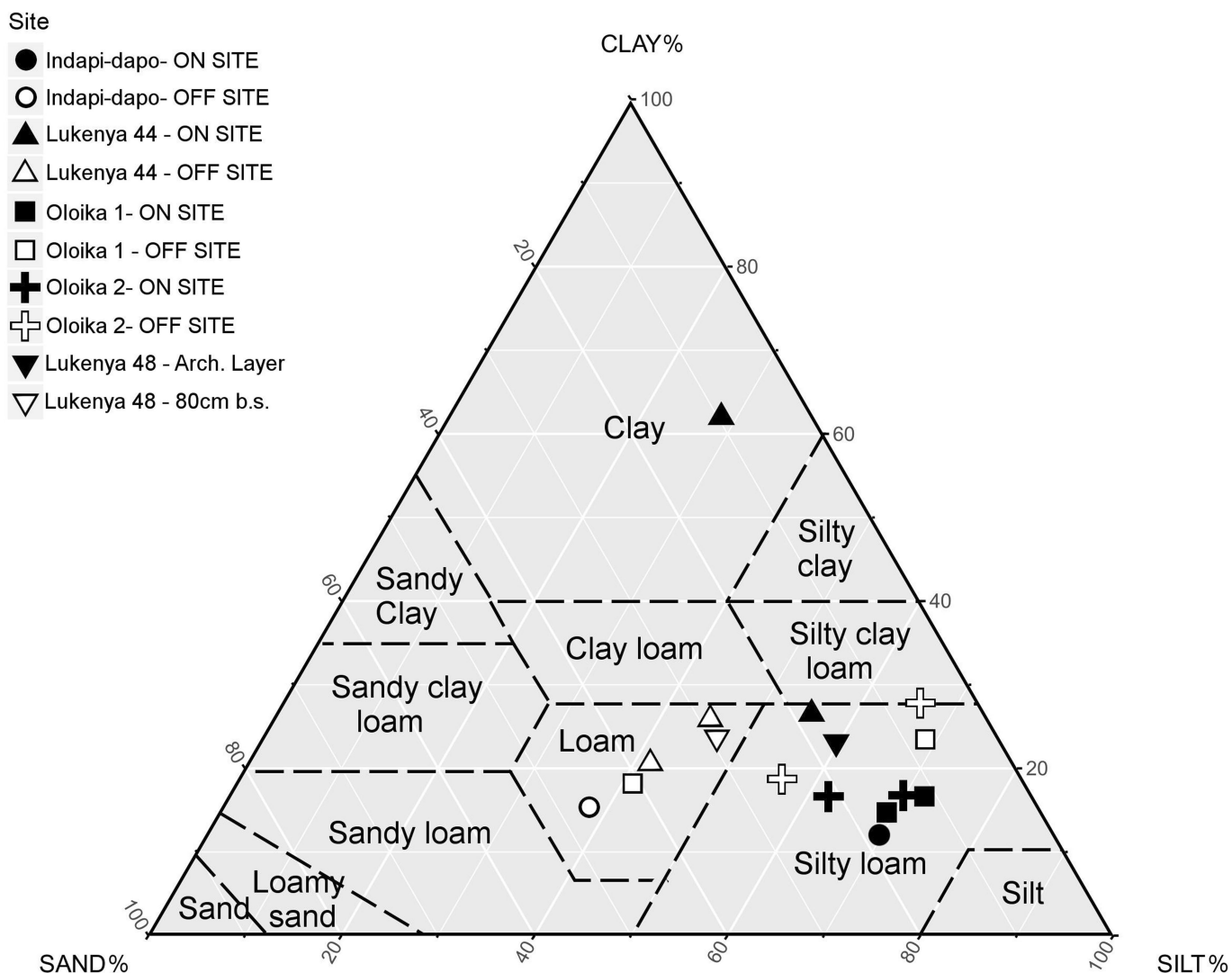
Extended Data Fig. 2 | Archaeological landscapes and stratigraphic sections. **a**, Satellite image of GvJm 44 and GvJm 48, Lukenya (dry season). At GvJm 48, a track exposes fine-grained grey midden deposits in an open grassy area. Redder sandy clays are exposed north and south of the site. **b**, Landscape and stratigraphic view of GvJm 44, showing dark Neolithic midden sediment in cross section. Arrows indicate midden edges. Person standing atop the centre of the midden is about 165-cm tall. **c**, Dung layer

at GvJm 48. **d**, Open glades visible near the Ntuka River (dry season) at Ol Owarukeri (GvJh 108), a large Elmenteitan (Pastoral Neolithic tradition dating to approximately 3,500–1,500 cal. BP) site with modern pastoralist settlement and two smaller Pastoral Neolithic sites, one with modern settlement. **a**, **d**, Imagery from Google Earth Pro, Digital Globe. **b**, **c**, Photographs by S.H.A., 1977–1978.

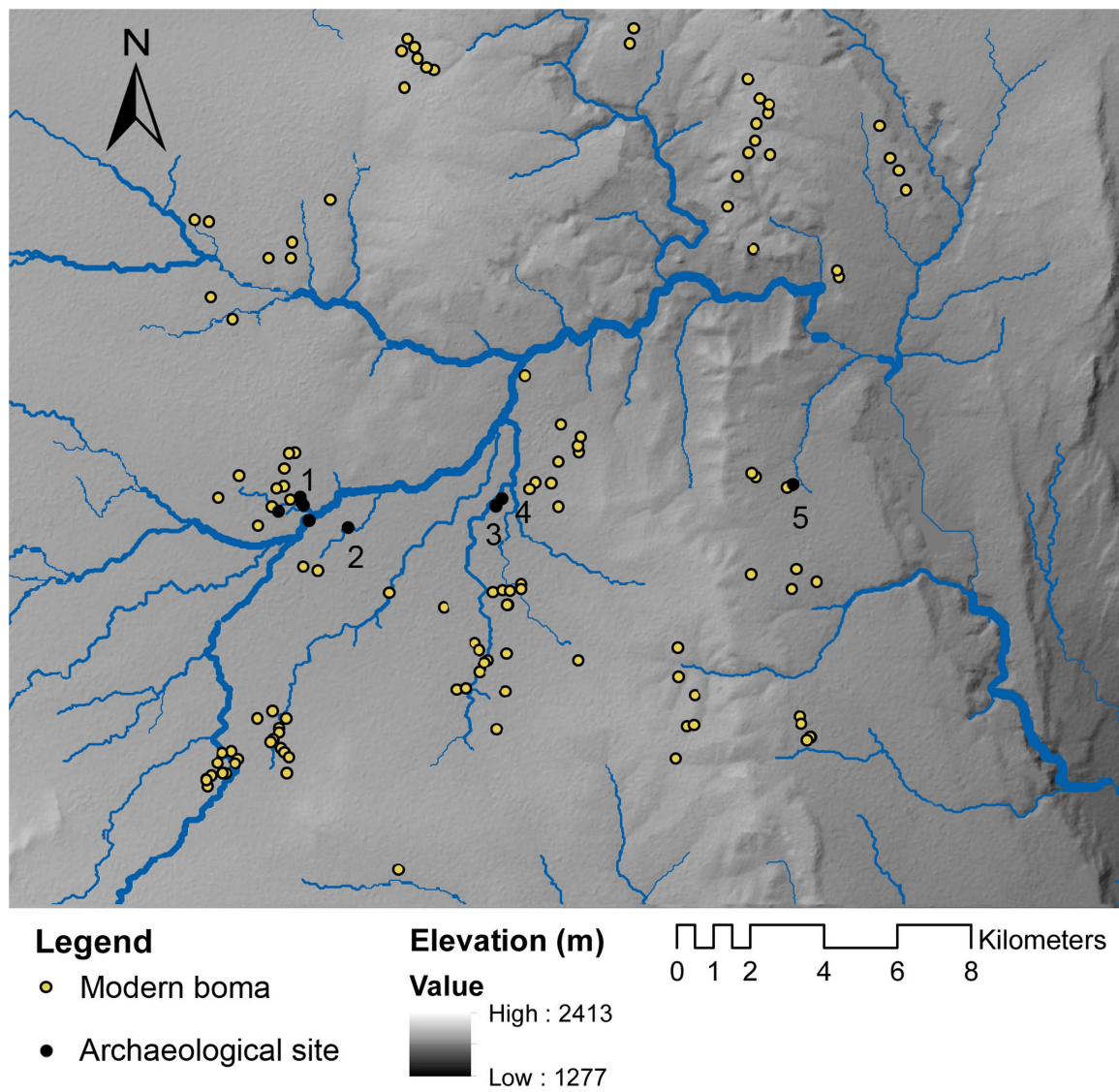


Extended Data Fig. 3 | Sediment sample micromorphology. **a**, Flatbed scan of a thin section representing off-site sediments (Oloika 1). **b**, Flatbed scan of a thin section representing on-site sediments (Oloika 1). Both scans are 6.2-cm wide. Note the colour and structure differences between on-site and off-site sediments. The reddish rounded particles are weathered local magmatic rock. **c**, Microphotograph of on-site sediments (Indapi Dapo) showing granular microstructure associated with large voids, which indicates severe bioturbation. Note the modern plant root (1) within the large void on the right. Scale bar, 1 mm; plane-polarized

light. **d**, Microphotograph of on-site sediments (Indapi Dapo) showing black manganese-oxide florets (2), which indicate periods of water saturation. Scale bar, 1 mm; plane-polarized light. **e**, Microphotograph of on-site sediments (Oloika 2) that have been disaggregated ('grain mount') to enable clear observation of phytoliths and dung spherulites. Arrows point to several phytoliths of various types. Scale bar, 0.1 mm; plane-polarized light. **f**, Same view as in **e**, but in crossed-polarized light. Arrows point to a few dung spherulites.



Extended Data Fig. 4 | Ternary plot of particle size distributions for sampled archaeological and off-site contexts. $n = 8$ archaeological contexts; $n = 8$ off-site contexts. See Supplementary Table 1 for values. Plot generated using the ggtern extension for ggplot2³³.



Extended Data Fig. 5 | Landscape of Nkuta showing 116 modern and 5 ancient pastoral settlements (bomas) visible in the study area. Ole Pariata (1), Ol Owarukeri (GvJh108) (2), Oloika 1 (3), Oloika 2 (4) and

Indapi Dapo (5). ArcGIS model with a base 30-m-resolution digital elevation derived from the Shuttle Radar Topography Mission.

Extended Data Table 1 | Archaeology

Site	Lat°/ Long°	Elevation (m)	Tradition	Glade area (m ²)	On-site PSA	Off-site PSA	On-site enriched	Off-site enriched	Carbonate layer
Oloika 1 (GvJh86)	S 1.36705 E 35.9709	1745	Elmenteitan	2320	Silty loam	More sand and clay	P, Mg, K [†]	Al	thick
Oloika 2 (GvJh85)	S 1.36545 E 35.9001	1745	Elmenteitan	4750	Silty loam	More clay	P, Mg [†]	Zr, Pb (Al)	thick
Indapi Dapo (GvJh121)	S 1.36176 E 35.9713	1650	Savannah Pastoral Neolithic	8360	Silty loam	More sand	P, Mg, Ca, Sr Na [†] , K [†]	Co, V, Ni, Zr, Pb	thin
Vaave Makongo (GvJm44)	S 1.4761 E 37.074	1686	Savannah Pastoral Neolithic	n.d.	Silty loam	More sand	P	n.d.	n.d.
Lukenya (GvJm48)	S 1.47266 E 37.0762	1683	Savannah Pastoral Neolithic	n.d.	Silty loam	More sand	Ca, P [†] , Mg [†] , Na [†]	n.d.	n.d.

n.d., no data. PSA, particle size analysis.

†Questionable or moderate enrichment.

Extended Data Table 2 | Radiocarbon dates

Site & material culture	Material	Lab #	C14 years	Calibrated years BP*	Reference
Indapi dapo (SPN, Narosura)	tooth dentine collagen	ISGS A3371	2420 ± 20	2461-2364	Reported here
Indapi dapo (SPN, Narosura)	tooth enamel apatite	ISGS A3372	2330 ± 15	2352-2342	Reported here
Oloika 1 GvJh85 (Elmenteitan)	charcoal	ISGS A2076	2420 ± 20	2461-2364	Reported here
Oloika 2 GvJh86 (Elmenteitan)	tooth enamel apatite	ISGS A2125	2095 ± 15	2113-2011	Reported here
Sugenya, (upper dung)	charcoal	Pta-9058	2230 ± 60	2312-2167	36
Sugenya, (lower dung)	charcoal	Pta-9063	2680 ± 60	2853-2764	36
GvJm44 Lukenya (SPN, Nderit)	charcoal	GX5348	3290 ± 145	3703-3361	37
GvJm44 Lukenya, (SPN, Narosura)	charcoal	GX5138	2415 ± 155	2714-2345	37
GvJm44 Lukenya, (SPN, Narosura)	bone apatite	GX4160-A	2085 ± 135	2301-1899	38
GvJm44 Lukenya, (SPN, Narosura)	bone collagen	GX4161-C	1710 ± 135	1813-1422	38
GvJm44 Lukenya, (SPN, Akira)	bone gelatin	GX5638-G	2070 +155	2305-1875	38
GvJm44 Lukenya, (SPN, Akira)	bone apatite	GX4507-A	2030 ± 125	2150-1830	38
GvJm44 Lukenya, (SPN, Akira)	bone apatite	GX5638-A	1820 ± 200	1990-1531	37
GvJm48 Lukenya (SPN, Narosura)	bone gelatin	GX5347-G	1810 ± 135	1879-1569	37
GvJm48 Lukenya, (SPN, Narosura)	bone apatite	GX5347-A	1600 ± 130	1685-1354	37

SPN, Savannah Pastoral Neolithic.

*68.2% confidence interval. Calibrated using SHCAL13 in OxCal 4.2³⁴⁻³⁸.

A cortical filter that learns to suppress the acoustic consequences of movement

David M. Schneider^{1,2,3}, Janani Sundararajan^{1,3} & Richard Mooney^{1*}

Sounds can arise from the environment and also predictably from many of our own movements, such as vocalizing, walking, or playing music. The capacity to anticipate these movement-related (reafferent) sounds and distinguish them from environmental sounds is essential for normal hearing^{1,2}, but the neural circuits that learn to anticipate the often arbitrary and changeable sounds that result from our movements remain largely unknown. Here we developed an acoustic virtual reality (aVR) system in which a mouse learned to associate a novel sound with its locomotor movements, allowing us to identify the neural circuit mechanisms that learn to suppress refferent sounds and to probe the behavioural consequences of this predictable sensorimotor experience. We found that aVR experience gradually and selectively suppressed auditory cortical responses to the refferent frequency, in part

by strengthening motor cortical activation of auditory cortical inhibitory neurons that respond to the refferent tone. This plasticity is behaviourally adaptive, as aVR-experienced mice showed an enhanced ability to detect non-refferent tones during movement. Together, these findings describe a dynamic sensory filter that involves motor cortical inputs to the auditory cortex that can be shaped by experience to selectively suppress the predictable acoustic consequences of movement.

Auditory activity in the brains of humans and other mammals is suppressed during a wide variety of movements, including vocalization and locomotion^{1,3–9}. Although the stereotyped and often simple acoustic consequences (that is, auditory refference) of rhythmic movements such as licking or chewing can be suppressed by brainstem mechanisms⁸, a more flexible form of movement-related suppression is

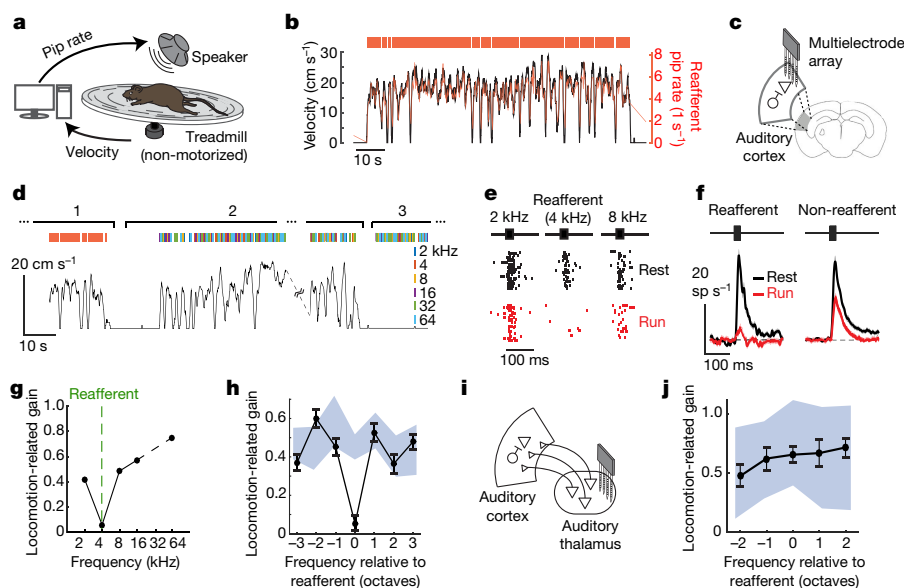


Fig. 1 | Locomotion-related suppression is specific for the frequency of self-generated sounds. **a**, Mice were acclimated to a closed-loop acoustic virtual reality (aVR) system. **b**, Example traces of locomotion (black trace), timing of individual aVR tone pips (red ticks), and instantaneous rate of aVR tone pips (red trace). **c**, Extracellular recordings were made from auditory cortical neurons during running and resting. The mouse brain in this figure has been reproduced with permission²⁷. **d**, During electrophysiology, mice first heard the refferent tone frequency that they expected treadmill running to produce (for example, 4 kHz) (phase 1). During phase 2, mice heard tones that were yoked in time to their running speed but with random frequency. During phase 3, mice heard tones of random frequency while at rest. **e**, Action potential responses from an example neuron while at rest (black, phase 3 from **d**) and while running (red, phase 2 from **d**) following playback of the expected refferent frequency (4 kHz) as well as tone frequencies one octave away. **f**, Population peri-stimulus-time histograms (PSTHs) showing neural

responses to refferent (left) and non-refferent (right) frequencies during running (red) and resting (black) conditions. Responses to non-refferent sounds are averaged across all five non-refferent frequencies. During running, responses to the expected refferent frequency were suppressed relative to non-refferent frequencies ($N = 11$ mice, $n = 317$ neurons, $P = 1.1 \times 10^{-18}$). **g**, Locomotion-related suppression of an example auditory cortical neuron. **h**, Average locomotion-related suppression (black, mean \pm s.e.) of auditory cortical neurons, centred on the expected refferent frequency heard by each mouse ($N = 11$ mice, $n = 317$ neurons). Blue area shows 95% confidence bounds. **i**, Extracellular recordings were made from the auditory thalamus during running and resting following 6–9 days of aVR acclimation. **j**, Locomotion-related suppression (black, mean \pm s.e.) of auditory thalamic neurons ($N = 5$ mice, $n = 109$ neurons) is not specific to the expected refferent frequency. For statistical details, see Methods.

¹Department of Neurobiology, Duke University School of Medicine, Durham, NC, USA. ²Center for Neural Science, New York University, New York, NY, USA. ³These authors contributed equally: David M. Schneider, Janani Sundararajan. *e-mail: mooney@neuro.duke.edu

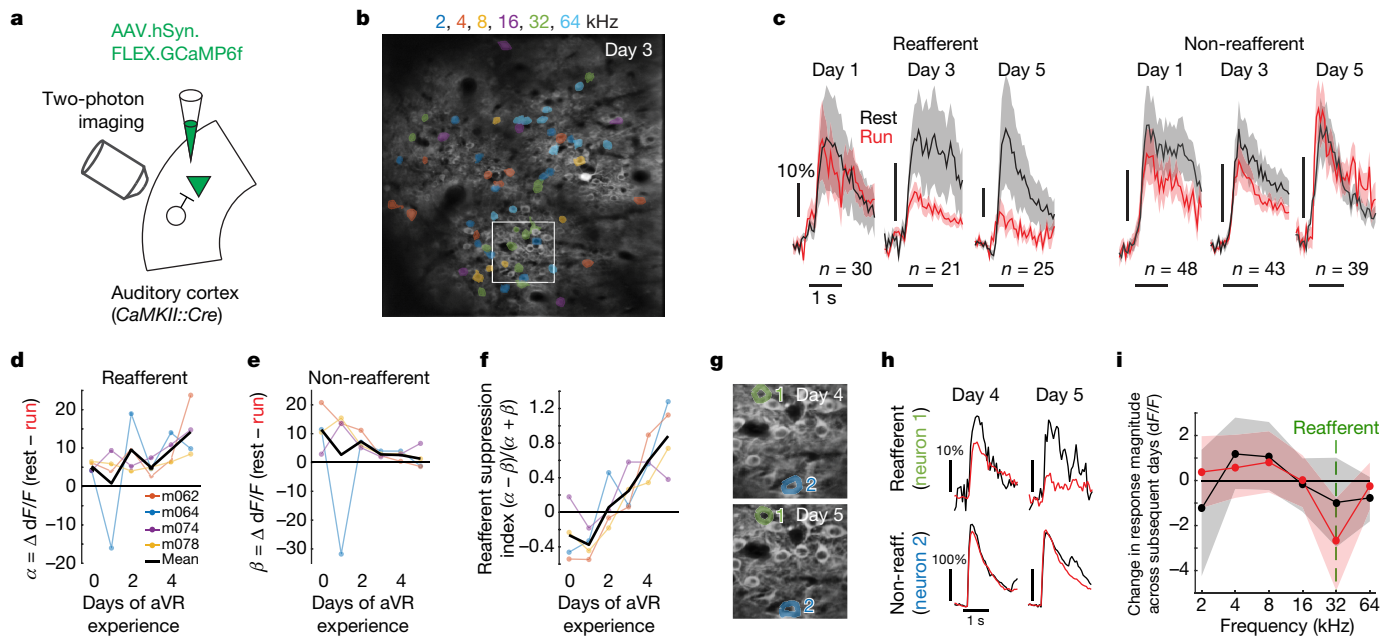


Fig. 2 | Reafferent suppression arises in parallel with sensory-motor experience. **a**, Injection of AAV encoding Cre-dependent GCaMP6f into auditory cortex of *CaMKII::Cre* mouse followed by calcium imaging of layer 2/3 excitatory neurons during aVR experience over several days. **b**, Maximum projection of an example field of layer 2/3 excitatory neurons in auditory cortex from a mouse acclimated to aVR experience producing 32-kHz tones. Colour represents the best frequency of every neuron. White box shows region in **g**. **c**, Population-averaged calcium transients evoked by the expected refferent frequency (left) and a non-refferent frequency (right, two octaves away) during rest (black) and running (red) following 1, 3 and 5 days of aVR experience. Shaded areas show mean \pm s.e. **d**, Magnitude of locomotion-related suppression (α) for the expected refferent frequency ($N = 4$ mice). **e**, Magnitude of locomotion-related suppression (β) for non-refferent frequencies (two octaves

thought to arise independently in the auditory cortex through mechanisms that are not well understood^{1,6,10–13}. To begin to identify these cortical mechanisms, we developed an acoustic virtual reality (aVR) system in which head-fixed mice on a treadmill heard a series of tones presented at a rate proportional to their running speed, simulating an experimentally adjustable yet predictable form of auditory reafference associated with locomotion (Fig. 1a, b; Extended Data Fig. 1a–d; Supplementary Video 1). After about a week of aVR experience, auditory cortical responses to the reafferent tone during locomotion were nearly abolished, whereas responses to frequencies one or more octaves distant from the reafferent frequency showed more modest suppression (similar to that seen during locomotion in aVR-naïve mice¹; Fig. 1c–f). Notably, this differential suppression manifested only during movement; in resting mice, a similar fraction of auditory cortical neurons responded to reafferent and non-refferent tones, and they did so with equivalent firing rates (Fig. 1e, f and Extended Data Fig. 2a–d).

To further characterize the movement-dependent filter formed by aVR experience, we compared the frequency tuning curves of auditory cortical neurons at running and at rest and calculated a locomotion-related gain function for each neuron (Fig. 1g). Regardless of a neuron's best frequency, locomotion-related suppression was greater for the reafferent tone than for tones one or more octaves away, with tones half an octave from the reafferent frequency showing intermediate suppression during locomotion (Fig. 1h; Extended Data Figs. 2e–h, 3a). The width of this 'notch' filter may be constrained by the tuning widths of auditory cortical inhibitory neurons¹⁴ and more broadly conforms to the idea that excitatory neurons in sensory cortex receive input from local inhibitory neurons tuned to both similar and dissimilar stimulus features^{15,16}. Movement-related suppression in auditory thalamic

lower). **f**, Selectivity of locomotion-related suppression for the expected reafferent frequency [$(\alpha - \beta)/(\alpha + \beta)$]. Values greater than zero indicate stronger locomotion-related suppression at the reafferent frequency; values less than zero indicate stronger locomotion-related suppression at non-refferent frequencies. **g**, Maximum projection on subsequent days, zoomed in on the region outlined in white panel in **b**. **h**, Average calcium traces from two example neurons shown in **g** during days 4 and 5 of aVR experience. Responses for neuron 1 are to 32-kHz tones (the expected reafferent frequency) and responses for neuron 2 are to 2-kHz tones (a non-refferent frequency). Black traces, rest; red traces, running. **i**, In two mice, 241 sound-responsive neurons were monitored across six pairs of subsequent days of aVR experience. Change in tuning curves across subsequent days (black, rest; red, running). Shaded regions are 95% confidence bounds. For statistical details, see Methods.

neurons remained flat across sound frequencies (Fig. 1i, j), as previously observed in aVR-naïve mice^{1,4,17}, and reafferent tones were more strongly suppressed in infragranular than supragranular auditory cortex (Extended Data Fig. 3b–d), indicating that circuits local to the auditory cortex are a likely source of the reafferent notch filter that arises following aVR experience.

The formation of this auditory cortical filter required a predictable and relatively prolonged association of movement with an ensuing sound. Mice acclimated for about a week on a treadmill in which fixed-frequency tones were presented only at rest did not display enhanced cortical suppression at the training frequency during rest or running (Extended Data Fig. 4a, b). Furthermore, mice acclimated for about a week on a treadmill in which tones were presented at a fixed tempo during locomotion regardless of running speed (that is, 'metronome'-experienced mice) showed no enhanced locomotion-related auditory cortical suppression at the training frequency (Extended Data Fig. 4c, d). Moreover, in aVR-experienced mice, enhanced suppression of the reafferent tone was evident as soon as they began to move on the treadmill, whereas stimulus-specific adaptation emerged more slowly and only after several tone presentations (that is, measured at the non-refferent frequency; Extended Data Fig. 4e, f). Finally, one hour of aVR experience (around 1,000–3,000 reafferent tones) was not sufficient to induce any enhancement of locomotion-related suppression at the reafferent frequency (Extended Data Fig. 4g).

To determine more precisely the time course over which this notch filter arises, we used two-photon calcium methods to longitudinally image layer 2/3 excitatory neurons in the auditory cortex of mice across their first 5 days of aVR experience (Fig. 2a, b). Across days, tuning curves measured during rest were relatively stable ($r = 0.53$), but at

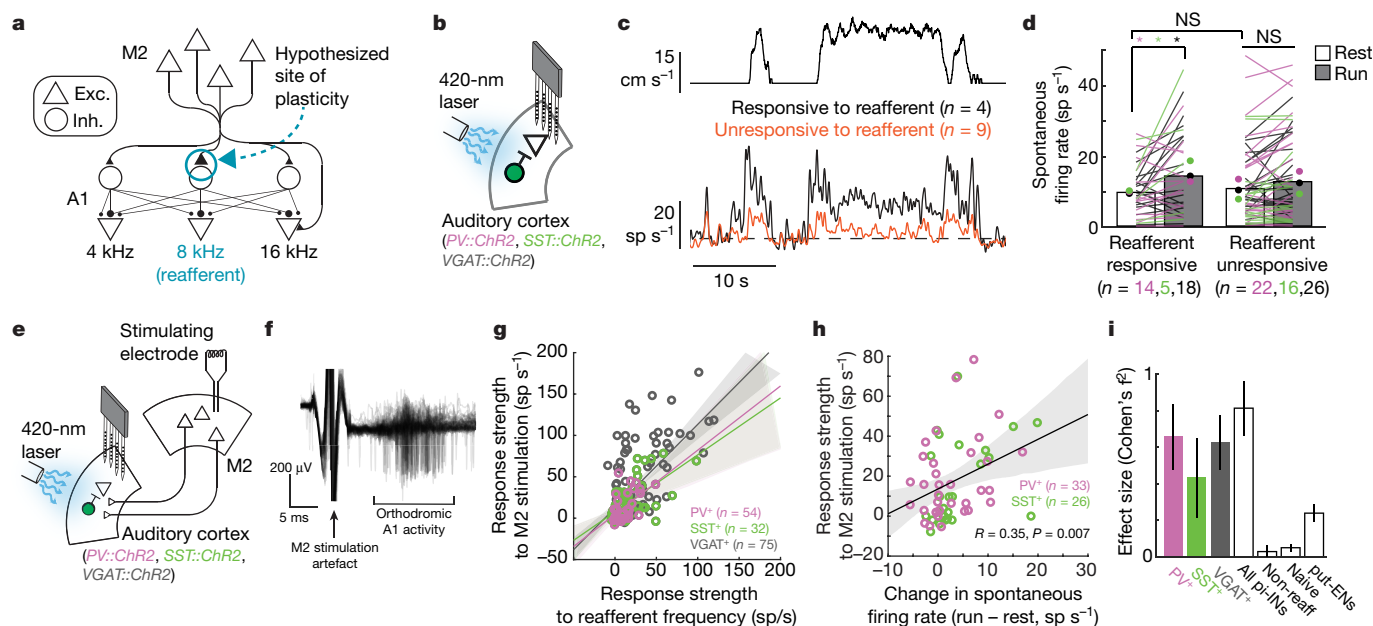


Fig. 3 | Reafferent-tuned inhibitory neurons increase their activity during locomotion and receive enhanced motor cortical input. **a**, Proposed model of experience-dependent strengthening of M2 inputs onto inhibitory neurons tuned to the reafferent frequency in the auditory cortex (blue circle). Synapse strength is proportional to size. **b**, Extracellular recording from photo-identified inhibitory neurons (pi-INs) in the auditory cortex. **c**, Top, locomotion velocity as a function of time. Bottom, spontaneous activity of simultaneously recorded pi-INs that were responsive to the reafferent frequency (black) or were not, but were responsive to other frequencies (orange). **d**, PV⁺ (magenta), SST⁺ (green) and VGAT⁺ (black) pi-INs that are responsive to the reafferent frequency increase their firing rate significantly during locomotion, unlike pi-INs that are not responsive to the expected reafferent frequency (* $P < 0.05$, Wilcoxon signed rank test). **e**, Extracellular recording from pi-INs in the auditory cortex. A bipolar microstimulating electrode was

implanted in M2. **f**, Example of M2 stimulation-evoked action potentials recorded from a VGAT⁺ pi-IN in auditory cortex. **g**, pi-INs that were more strongly driven by the reafferent frequency were more strongly recruited by electrical stimulation in M2. Solid lines and shadows show linear regression and 95% confidence bounds, respectively. **h**, Changes in spontaneous firing rate during locomotion relative to rest for pi-INs (PV⁺ and SST⁺) were significantly correlated with the magnitude of the same neurons' responses to M2 electrical stimulation. **i**, Effect size of the correlations shown in **g** and for four other conditions. All pi-INs: pi-INs from PV⁺, SST⁺ and VGAT⁺ mice. Non-reaff: same neurons as shown in **g** (VGAT⁺) but for responses to a non-reafferent tone. Naive: pi-INs from VGAT⁺ mice that were acclimated to a quiet treadmill. put-ENs: putative excitatory neurons in auditory cortex of VGAT⁺:ChR2 mice acclimated to aVR. Error bars show 95% confidence bounds. For statistical details, see Methods.

the population level, locomotion-related suppression became progressively more specific for the reafferent frequency (Fig. 2c). The emergence of frequency-specific suppression during locomotion involved both increased suppression of the reafferent frequency and decreased suppression of non-reafferent frequencies (Fig. 2d–f). Tracking the activity of a subset of the same neurons ($n = 241$ from two mice) across consecutive days of aVR experience revealed that their tuning curves measured during locomotion—but not during rest—changed only at the reafferent frequency, with individual neurons becoming less responsive across consecutive days (Fig. 2g–i).

Auditory cortical interneurons integrate auditory inputs with locomotor-related signals from the secondary motor cortex (M2)^{1,18,19} and form inhibitory synapses on both similarly and dissimilarly tuned excitatory cells, affording a substrate on which aVR experience could act to sculpt the movement-dependent notch filter described here^{1,18} (Fig. 3a). To explore this possibility, we expressed channelrhodopsin (ChR2) in genetically identified inhibitory neurons (parvalbumin (PV)⁺ (encoded by *PV* (also known as *Pvalb*)), somatostatin (SST)⁺, or vesicular γ -aminobutyric acid (GABA) transporter (VGAT)⁺ (encoded by *VGAT* (also known as *Slc32a1*)) neurons; see Methods), acclimated mice to aVR experience for about a week, and then recorded the action potential activity of photo-identified auditory cortical inhibitory cells (Fig. 3b and Extended Data Fig. 5a, b). During running, the spontaneous firing rates in only those inhibitory neurons responsive to the reafferent frequency increased, while their responses to reafferent and non-reafferent tones decreased modestly (Fig. 3c, d; Extended Data Fig. 5c). By contrast, the spontaneous firing rates of putative excitatory neurons decreased slightly during locomotion (Extended Data Fig. 5d). Moreover, the

magnitude of the running-related increase in an inhibitory neuron's spontaneous firing rate scaled with the strength of its response to the reafferent frequency (Extended Data Fig. 5e). Finally, locomotion-related suppression at the reafferent frequency comprised both divisive and subtractive components, consistent with the involvement of both PV⁺ and SST⁺ interneurons²⁰ (Extended Data Fig. 5f). Therefore, aVR experience selectively enhanced the movement-dependent recruitment of inhibitory interneurons that respond to the reafferent frequency.

One idea is that aVR experience strengthens the connections between M2 and auditory cortical interneurons that respond to the reafferent frequency, resulting in their selective recruitment during locomotion. To test this idea, we calculated frequency-tuning curves of photo-identified PV⁺, SST⁺ and VGAT⁺ inhibitory neurons in the auditory cortex of aVR-experienced mice. We then applied brief current pulses in M2 and measured the resulting action potential activity in these identified interneurons (Fig. 3e, f). Auditory cortical inhibitory neurons that responded strongly to the reafferent frequency were driven more strongly by M2 stimulation than were inhibitory neurons that responded only weakly or not at all to the reafferent frequency (Fig. 3g, i and Extended Data Fig. 5g, j). Moreover, the PV⁺ and SST⁺ inhibitory neurons that were activated most strongly by M2 stimulation in resting mice showed the greatest increases in spontaneous firing rates when the mouse was running (Fig. 3h). Using a similar approach, we also detected a similar but weaker correlation for auditory cortical excitatory neurons in aVR-experienced mice (Fig. 3i and Extended Data Fig. 5h, j). These differential effects of M2 stimulation on auditory cortical excitatory and inhibitory neurons are consistent with the observation that M2 synapses excite both cell types but exert a primarily suppressive effect on auditory cortical activity through feedforward

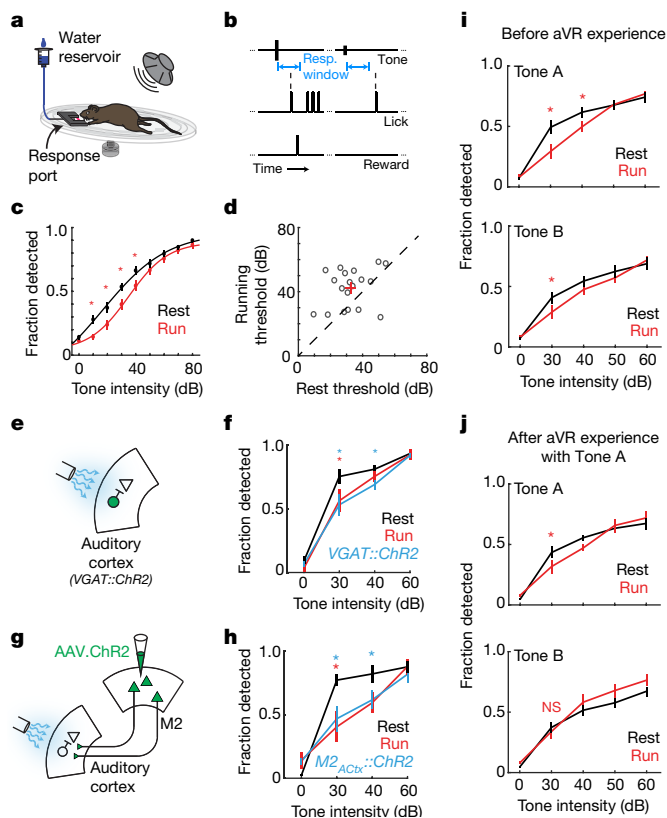


Fig. 4 | Tone detection behaviour is compromised by locomotion, is auditory-cortex dependent, and adapts following aVR experience. **a**, Mice were trained to lick a response port upon hearing a tone while resting or running on a treadmill. **b**, Tones were presented with random inter-tone intervals following a period without spontaneous licking. Mice were rewarded for licking within 1 s of tone presentation. **c**, Average psychometric functions showing detection rates as a function of tone intensity while mice were resting (black) or running (red). $N = 19$ mice, Red asterisk, $P < 0.005$. **d**, Behavioural threshold (intensity at 50% performance) during resting and running for each mouse (black circles; red plus denotes mean threshold). Dashed line shows unity. $N = 19$ mice, $P = 0.009$, paired t -test. **e**, Optogenetic activation of inhibitory neurons in auditory cortex during tone detection. **f**, Tone detection performance ($N = 4$ mice) during rest (black), running (red) and rest with optogenetic activation of auditory cortical inhibitory neurons (blue). Mice were worse at detecting tones during optogenetic trials compared to non-optogenetic trials at rest (blue asterisk, $P < 0.05$) and during running compared to rest (red asterisk, $P < 0.05$). **g**, Optogenetic activation of ChR2⁺ M2 terminals in auditory cortex during tone detection. **h**, Tone detection performance ($N = 4$ mice) during rest (black), running (red) and rest with optogenetic activation of M2 terminals in auditory cortex (blue). Mice were worse at detecting tones during optogenetic trials compared to non-optogenetic trials at rest (blue asterisk, $P < 0.05$) and during running compared to rest (red asterisk, $P < 0.05$). **i**, Tone detection performance ($N = 10$ mice) during rest and running for naive mice performing a two-frequency detection task (red asterisk, $P < 0.05$). **j**, As in **i**, but following aVR experience with Tone A (red asterisk, $P < 0.05$). Error bars show s.e.m. For further statistical details, see Methods and Supplementary Table 1.

inhibition^{1,18}. Finally, in aVR-naive mice, there was only a weak correlation between tone-evoked and M2-stimulation-evoked responses in auditory cortical interneurons (Fig. 3i and Extended Data Fig. 5i, j).

To investigate how movement-related signals in the auditory cortex influence auditory perception, we trained mice to detect tones while running and resting and confirmed that this tone detection task required auditory cortical activity (Fig. 4a, b; Supplementary Video 2; Extended Data Fig. 6a). This test revealed that mice were significantly worse at detecting tones at intermediate intensities (10–40 dB) while running than at rest, despite displaying equal levels of motivation in

these two states (Fig. 4c, d; Extended Data Fig. 6b). Furthermore, optogenetic activation of inhibitory interneurons or M2 axon terminals in the auditory cortex degraded performance in resting mice at these intermediate sound intensities, indicating that the same circuit elements that are influenced by aVR experience modulate hearing in a movement-dependent manner (Fig. 4e–h; Extended Data Fig. 6c–f). By contrast, optogenetically activating inhibitory interneurons in visual cortex, illuminating GFP-expressing M2 terminals in the auditory cortex, or simply illuminating the intact skull in resting mice did not diminish their ability to detect tones (Extended Data Fig. 6g–i).

A remaining issue is whether the auditory cortical filter formed by aVR experience improves the mouse's ability to detect non-reafferent tones during movement, as hypothesized for brain mechanisms that suppress predictable sensory reafference^{2,21,22}. We trained mice to detect two different tones (tones A and B, separated by two octaves), which were randomly interleaved from one trial to the next. After several days of training, mice showed a similar deficit in detecting both tones during locomotion, similar to mice trained on a single tone (Fig. 4i). Mice then received about a week of aVR experience in which only one of the tones (tone A) was used as a reafferent training stimulus. Following aVR experience, mice no longer showed a locomotion-related deficit in detecting the non-reafferent tone (tone B), even though they continued to show a movement-related deficit in detecting the reafferent tone (tone A) (Fig. 4i, j). Therefore, aVR experience not only selectively suppresses auditory cortical responses to predictable reafferent sounds, but also improves the ability of mice to detect unpredictable sounds during movement (Extended Data Fig. 6j, k).

Our findings establish that temporally coupled locomotor–auditory experience results in the formation of a movement-dependent filter that suppresses auditory cortical responses to predictable self-generated sounds. A plausible idea is that coincident motor and auditory activity during sound-generating movements strengthens M2 synapses onto PV⁺ and SST⁺ interneurons, or onto neurons interposed between M2 and these auditory cortical interneurons, leading to enhanced movement-related suppression of auditory cortical responses to the reafferent tone²³ and an enhanced ability to detect non-reafferent tones during movement^{24,25}. Notably, aVR experience also reduced locomotion-dependent suppression at non-reafferent frequencies in layer 2/3 of the auditory cortex, providing an auditory cortical correlate of this adaptive perceptual change. The involvement of M2 in this form of auditory cortical suppression is reminiscent of the motor cortex-mediated suppression of responses to predictable stimuli in mouse primary visual cortex²⁶, consistent with a generalized predictive cortical mechanism. Ultimately, the motor–auditory cortical circuit characterized here can flexibly encode the relationship between a movement and the sound it produces, helping to maintain sensitivity to novel sounds in the environment while also monitoring the acoustic consequences of sound-generating movements.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0520-5>.

Received: 31 July 2017; Accepted: 23 July 2018;
Published online 12 September 2018.

- Schneider, D. M., Nelson, A. & Mooney, R. A synaptic and circuit basis for corollary discharge in the auditory cortex. *Nature* **513**, 189–194 (2014).
- Weiss, C., Herwig, A. & Schütz-Bosbach, S. The self in action effects: selective attenuation of self-generated sounds. *Cognition* **121**, 207–218 (2011).
- Kuchibhotla, K. V. et al. Parallel processing by cortical inhibition enables context-dependent behavior. *Nat. Neurosci.* **20**, 62–71 (2017).
- Zhou, M. et al. Scaling down of balanced excitation and inhibition by active behavioral states in auditory cortex. *Nat. Neurosci.* **17**, 841–850 (2014).
- Rummell, B. P., Klee, J. L. & Sigurdsson, T. Attenuation of responses to self-generated sounds in auditory cortical neurons. *J. Neurosci.* **36**, 12010–12026 (2016).
- Flinker, A. et al. Single-trial speech suppression of auditory cortex activity in humans. *J. Neurosci.* **30**, 16643–16650 (2010).

7. Eliades, S. J. & Wang, X. Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations. *J. Neurophysiol.* **89**, 2194–2207 (2003).
8. Singla, S., Dempsey, C., Warren, R., Enikolopov, A. G. & Sawtell, N. B. A cerebellum-like circuit in the auditory system cancels responses to self-generated sounds. *Nat. Neurosci.* **20**, 943–950 (2017).
9. Curio, G., Neuloh, G., Numminen, J., Jousmäki, V. & Hari, R. Speaking modifies voice-evoked activity in the human auditory cortex. *Hum. Brain Mapp.* **9**, 183–191 (2000).
10. Keller, G. B. & Hahnloser, R. H. R. Neural processing of auditory feedback during vocal practice in a songbird. *Nature* **457**, 187–190 (2009).
11. Eliades, S. J. & Wang, X. Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature* **453**, 1102–1106 (2008).
12. Houde, J. F. & Jordan, M. I. Sensorimotor adaptation in speech production. *Science* **279**, 1213–1216 (1998).
13. Mifsud, N. G. & Whitford, T. J. Sensory attenuation of self-initiated sounds maps onto habitual associations between motor action and sound. *Neuropsychologia* **103**, 38–43 (2017).
14. Moore, A. K. & Wehr, M. Parvalbumin-expressing inhibitory interneurons in auditory cortex are well-tuned for frequency. *J. Neurosci.* **33**, 13713–13723 (2013).
15. Fino, E. & Yuste, R. Dense inhibitory connectivity in neocortex. *Neuron* **69**, 1188–1203 (2011).
16. Znamenskiy, P. et al. Functional selectivity and specific connectivity of inhibitory neurons in primary visual cortex. Preprint at <https://www.biorxiv.org/content/early/2018/04/04/294835> (2018).
17. Williamson, R. S., Hancock, K. E., Shinn-Cunningham, B. G. & Polley, D. B. Locomotion and task demands differentially modulate thalamic audiovisual processing during active search. *Curr. Biol.* **25**, 1885–1891 (2015).
18. Nelson, A. et al. A circuit for motor cortical modulation of auditory cortical activity. *J. Neurosci.* **33**, 14342–14353 (2013).
19. Nelson, A. & Mooney, R. The basal forebrain and motor cortex provide convergent yet distinct movement-related inputs to the auditory cortex. *Neuron* **90**, 635–648 (2016).
20. Wilson, N. R., Runyan, C. A., Wang, F. L. & Sur, M. Division and subtraction by distinct cortical inhibitory networks *in vivo*. *Nature* **488**, 343–348 (2012).
21. Wolpert, D. M., Ghahramani, Z. & Jordan, M. I. An internal model for sensorimotor integration. *Science* **269**, 1880–1882 (1995).
22. Keller, G. B., Bonhoeffer, T. & Hübner, M. Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron* **74**, 809–815 (2012).
23. Froemke, R. C., Merzenich, M. M. & Schreiner, C. E. A synaptic memory trace for cortical receptive field plasticity. *Nature* **450**, 425–429 (2007).
24. Froemke, R. C. et al. Long-term modification of cortical synapses improves sensory perception. *Nat. Neurosci.* **16**, 79–88 (2013).
25. McGinley, M. J., David, S. V. & McCormick, D. A. Cortical membrane potential signature of optimal states for sensory signal detection. *Neuron* **87**, 179–192 (2015).
26. Leinweber, M., Ward, D. R., Sobczak, J. M., Attinger, A. & Keller, G. B. A sensorimotor circuit in mouse cortex for visual flow predictions. *Neuron* **95**, 1420–1432.e5 (2017).
27. Franklin, K. B. & Paxinos, G. *The Mouse Brain in Stereotaxic Coordinates, Compact The Coronal Plates and Diagrams* (Elsevier, Amsterdam, 2008).

Acknowledgements We thank K. Tschida, M. Tanaka, and D. Purves for their comments on this manuscript; members of the Mooney laboratory for discussions regarding experimental design and data analysis; J. Pearson for comments regarding statistical analyses; and M. Booze for animal care and technical support. This research was supported by an HHMI fellowship of the Helen Hay Whitney Foundation and a Career Award at the Scientific Interface from the Burroughs Wellcome Fund (D.M.S.), the Holland-Trice Graduate Fellowship in Brain Sciences (J.S.), and NIH grant 5 R01 DC013826 (R.M.).

Reviewer information *Nature* thanks S. Eliades, G. Keller and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions D.M.S., J.S., and R.M. initiated the project and designed the experiments. D.M.S. designed the aVR system, performed electrophysiology, optogenetic and calcium imaging experiments, and helped to design the psychophysics platform. J.S. designed the psychophysics platform, and performed psychophysics, pharmacology and optogenetic behavioural experiments. D.M.S. and J.S. analysed the data. D.M.S., J.S. and R.M. prepared the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0520-5>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0520-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Surgical procedures. All experimental protocols were approved by Duke University Institutional Animal Care and Use Committee. Male and female mice (*Mus musculus*) were purchased from Jackson Labs and were housed and bred in an onsite vivarium. We used mice that were 2–4 months old for our experiments. During all experiments, mice were kept on a reverse day–night cycle (12 h day, 12 h night).

For all surgical procedures, mice were anaesthetized under isoflurane (1–2% in O₂) and placed in a stereotaxic holder (Leica), skin was removed over the top of the head, and a titanium headpost was attached to the skull using a transparent adhesive (Metabond). Prior to electrophysiology experiments and following aVR acclimation, in male and female mice (C57, *VGAT::ChR2*, *PV::Cre*, *SST::Cre*), craniotomies were made to expose auditory cortex, auditory thalamus, and/or motor cortex to allow electrophysiology or electrical stimulation. A small craniotomy was made over the right sensory cortex and a silver pellet was positioned atop the cortical surface and cemented in place (Metabond) for use as a ground electrode. Exposed craniotomies were covered with a silicone elastomer (Kwik-Sil) and the mouse was allowed to recover in its home cage.

Prior to calcium imaging experiments, an AAV encoding Cre-dependent GCaMP6f was injected into the auditory cortex of *CaMKII::Cre* mice (see Viral injections), a tattoo was placed on the surface of the skull to mark the injection site, and a custom Y-shaped titanium headpost was attached to the skull with Metabond. Exposed skull on the top and side was also covered in Metabond. Mice were allowed to recover in their home cage for three weeks, after which time mice were again anaesthetized and a rectangular craniotomy was made over the original injection site. A stack of two laminated glass coverslips was placed over the craniotomy and sealed with Metabond. Mice were returned to their home cage and allowed to recover for 1 to 2 days.

For simultaneous psychophysics and pharmacological manipulations or optogenetic stimulation experiments (M2 terminals or *VGAT*⁺ neurons) in the auditory cortex, a custom Y-shaped titanium headpost was attached to the skull with Metabond and marks were bilaterally placed on the surface of the skull over the auditory cortex. For experiments in which M2 axon terminals were activated optogenetically, an AAV encoding either Channelrhodopsin or eGFP was injected bilaterally into M2 at the same time as the headpost implant (see Viral injections). Once mice were proficient with the task, they were anaesthetized again under isoflurane and craniotomies were opened bilaterally over the auditory cortex, which was confirmed by recordings (Carbostar-1, Kation Scientific) performed at multiple locations within the craniotomy in anaesthetized mice to ensure auditory responses to 8-kHz tones (1–2 s inter-tone interval). For pharmacological experiments, exposed craniotomies were covered with a silicone elastomer (Kwik-Sil) and mice were allowed to recover in their home cage for a day before behavioural testing. For optogenetic stimulation experiments, circular glass coverslips (3 mm diameter) were implanted bilaterally over the auditory cortex and were sealed in place with Metabond. For simultaneous psychophysics and optogenetic manipulation experiments in the visual cortex (*VGAT*⁺), circular glass coverslips were bilaterally implanted over visual cortex (identified using stereotaxic coordinates) at the same time as an annular headpost was attached to the skull. Mice were returned to their home cage and allowed to recover for 1 to 2 days after implantation.

Viral injections. For expression of calcium indicators, the skull over the auditory cortex of male and female *CaMKII::Cre* mice was exposed and two small craniotomies were made over the auditory cortical surface (estimated using stereotaxic coordinates) separated by approximately 300 μ m along the rostral–caudal dimension. A pipette was backfilled with AAV1.hSyn.FLEX.GCaMP6f, angled at 30° relative to vertical, and lowered into the auditory cortex. Approximately 150–200 nl virus was pressure injected (Nanoject) into the centre of auditory cortex over the course of 15 min, repeated for each craniotomy. For expression of Channelrhodopsin or eGFP, the skull over M2 of C57, or auditory cortex for *PV::Cre* and *SST::Cre* male and female mice, was exposed. A single craniotomy was made over M2 or auditory cortex (stereotaxic coordinates), and approximately 300 nl AAV1.hSyn.ChR2.EYFP.WPRE (M2), AAV1.CB7.eGFP.WPRE (M2) or AAV1.EF1 α .DIO.ChR2.EYFP.WPRE (A1) was pressure injected over the course of 20 min. Following injections, craniotomies were filled with melted bone wax and the injection sites were covered in Metabond.

Acoustic virtual reality. We designed a custom acoustic virtual reality (aVR) system for yoking a series of fixed-frequency tone pips (25 ms with 5 ms cosine ramp onset and offset) to a mouse's running speed. To create the aVR system, we built a non-motorized treadmill from a 6-inch Plexiglas disk (Delvies Plastic) that was coated with a thin silicone sheet (Durometer, Marian Chicago) mounted to the post of a rotary encoder (US Digital). Output from the rotary encoder was monitored with a data acquisition card (National Instruments) connected to a computer (Dell) running custom Matlab software (Mathworks, PsychToolBox) and sampled at ~30 Hz. The computer was also connected to a sound card (RME Fireface UCX), the output of which was routed to an ultrasonic speaker (Tucker Davis

Technologies) located lateral to the mouse, ~15 cm from the mouse's right ear. We recorded the noise produced by the mouse's footsteps on the treadmill and the sound of the rotating treadmill itself during running by placing an ultrasonic microphone close (~1 cm) to the mouse's ear. We measured <1 dB increase (estimated by taking the root mean square (r.m.s.) value of 5-s segments of recordings) in the noise produced when the mouse was running on the treadmill compared to rest.

To calculate the inter-tone interval, we computed a filtered version of the mouse's velocity, which was the median of the last five velocity samples. Upon every sampling period, the desired inter-tone interval was updated to be proportional to the reciprocal of the current median-filtered velocity, scaled such that the rate of tone presentations closely matched the foot step rate, which was calculated from videos of mice running on the treadmill at various speeds. At speeds greater than 30 cm/s, the inter-tone-interval saturated at 100 ms to ensure spacing between tones. For the anti-coupled version of aVR, sounds were not presented while mice were running but the total number of tones that should have been presented, and the calculated intervals between them, were stored in memory. During rest, tones were played back to the mouse with inter-tone intervals drawn from the intervals that the mouse would have heard while running until the number of resting tones equalled the number of tones that mouse should have heard while running. For the metronome aVR, sounds were presented only during running and at a fixed rate (~2/s) that was not modulated by running speed.

Mice were held in place using two clamps (Altos Photonics) that secured the arms of the headpost (see Surgical procedures). On their first day of treadmill experience, the aVR system was turned off and mice ran and rested for 2 h without hearing any tones. Beginning on the second day (referred to as day 1 of aVR experience), tones of a fixed frequency were yoked to the mouse's velocity as described above. For each mouse, the tone frequency was fixed during the first 6–9 days of aVR experience at 2, 4, 8, 16, 32, or 64 kHz. Mice were placed on the treadmill for ~2 h per day and were free to transition between periods of running and rest, which typically occurred several times during each 2-h aVR acclimation session.

Electrophysiology and aVR. Following 6–9 days of aVR acclimation, mice were positioned atop the treadmill and a 32-channel electrode (Neuronexus, 4 × 8 configuration) was implanted into the auditory cortex or auditory thalamus. The electrode was connected to a digitizing headstage (Intan) and electrode signals were acquired, monitored in real time, and stored for subsequent offline analysis (OpenEphys). The electrode was allowed to settle for ~30 min, during which time mice ran on the treadmill and heard tones of the reafferent frequency to which they had been acclimated. Following this initial 30 min, the frequency of the running-related tones was switched from a fixed frequency to a pseudo-random distribution comprising 2, 4, 8, 16, 32 and 64 kHz tones. In a subset of mice ($N = 4$) we also included tones spaced half an octave higher and lower than the reafferent frequency. Random-frequency tones were presented with inter-tone intervals dictated by running speed. Random-frequency tones yoked to running continued until the mouse heard 50 to 100 tones of each frequency ($n = 7$ mice, 2 to 15 min) or for 30 min ($n = 4$ mice). After this time, tones with random frequency were presented during a period of rest with inter-tone intervals drawn from the distribution that the mouse had recently heard while running. Electrode signals were filtered (300 to 5,000 Hz) and action potentials from individual neurons were sorted offline for each electrode independently based on visualization of the action potential waveform and principal component analysis of the waveform using custom Matlab software (PostHawk, D.M.S.).

Tone-evoked action potential responses were measured for each neuron at each tone frequency, independently for tones presented during running and during rest. To calculate population PSTHs, the tone-evoked response of every neuron that was responsive to a particular frequency (see Statistical Methods) was averaged independently for running and resting conditions. For each neuron we measured the response strength to each tone frequency ($RS(f)$) as the firing rate following tone presentation minus the baseline firing rate. To calculate locomotion-related gain, we took the ratio of $RS(f)$ measured during rest and running ($gain(f) = RS(f)_{run}/RS(f)_{rest}$). To average the locomotion-related gain functions across mice acclimated to aVR producing reafferent tones of different frequencies, we aligned the gain function for each neuron to the reafferent frequency experienced by the mouse from which the neuron was recorded.

Recordings from photo-identified inhibitory neurons (pi-INs) were made in *VGAT::ChR2* mice (Jackson labs) and in *PV::Cre* and *SST::Cre* mice (Jackson labs) injected with Cre-dependent ChR2 (see Viral injections). During electrophysiology, a multi-electrode array was implanted in the auditory cortex and an optical fibre coupled to a blue laser (420 nm, Shanghai) was directed at the auditory cortical surface. Action potential responses were analysed in response to a series of 30 laser pulses (100 ms each, separated by 1 s, laser power: 15–30 mW). pi-INs were identified based on short-latency, high-reliability responses to optical stimulation (Extended Data Fig. 5). Neurons that were not classified as pi-INs in *VGAT::ChR2* mice were classified as putative excitatory neurons (put-ENs). Following aVR

experience, we presented tones while recording from pi-INs during rest to measure their frequency tuning curves. The best frequency of each neuron was computed as the frequency that drove the strongest response and was restricted to neurons that were significantly driven by tones of at least one frequency. We recorded the spontaneous activity of pi-INs as mice transitioned between periods of running and rest.

To measure the strength of the functional connection between M2 and auditory cortex neurons, we implanted a bipolar stimulating electrode into M2 and we electrically stimulated within M2 (100 μ s, 300 μ A) every 2 s while recording from pi-INs and put-ENs in auditory cortex. Response strength to M2 stimulation was calculated as the difference between the firing rate in a baseline window immediately preceding electrical stimulation (100-ms window) and in a brief window 13 to 40 ms after electrical stimulation. We performed a bootstrap analysis 1,000 times to compute the 95% confidence bounds of the regression lines. Effect size was computed as Cohen's f^2 , which is defined as $R^2/(1-R^2)$, where R is the Pearson correlation.

To measure locomotion-related suppression across cortical layers, in a subset of mice ($N=3$) we implanted the electrode perpendicular to the surface of the auditory cortical surface. To estimate the cortical layer in which each neuron on the 4×8 electrode resided, we calculated the current-source density (CSD) triggered off of tone playback. First, we created a two-dimensional map of local-field potential (LFP, bandpass filtered to include 0.1 to 70 Hz) activity by averaging across electrode shanks that were at the same depth, and plotting the averaged activity at each depth as a function of time relative to tone onset. We then computed the CSD as the second spatial derivative of the depth-specific LFP signal. We estimated which electrodes were in cortical layer 4 based on transitions between sources and sinks in the CSD and based on latency of tone-evoked responses. We then separated our electrodes into those residing in infragranular (deeper than layer 4) and supragranular (superficial to layer 4) layers of the cortex. We subsequently analysed the tone-evoked LFP signal to measure locomotion-related suppression⁴.

Calcium imaging and aVR. Three weeks after viral infection with GCaMP6f and one day after the implantation of a cranial window, mice were positioned atop the treadmill and under a resonant scanning two-photon microscope (Neurolabware) with a mode-locked titanium sapphire laser (Mai Tai DeepSee) at 920 nm (laser power levels: 50–130 mW). The microscope objective (16 \times 0.8 NA water immersion, Nikon) was angled at $\sim 35^\circ$ such that imaging was performed perpendicular to the auditory cortical surface while mice were positioned on the treadmill in a normal, upright position. Prior to aVR acclimation, we monitored the locomotion-related suppression of aVR-naïve mice independently at each frequency. We then chose as the reafferent frequency on subsequent days the sound frequency that had the least amount of movement-related suppression for neurons in our field of view. By using this approach, we did not bias ourselves towards sound frequencies that were already strongly suppressed. On each imaging day, mice ran on the aVR treadmill and heard fixed-frequency tones yoked to their running speed for 2 h, during which time images were not acquired. Following each 2 h aVR session, tones of random frequency (2 to 64 kHz) were presented during running (with timing yoked to the mouse's running speed) and rest (with timing chosen from the running inter-tone-intervals) during image acquisition at 15.5 Hz until mice heard at least 50 tones of each frequency, typically lasting for approximately 5 to 10 min.

GCaMP6f fluorescence images were registered to correct for movement artefact in the horizontal plane. Regions of interest (ROIs) were selected using a semi-automated identification method based on nearby correlated pixel activity (Scanbox) and by manually tracing around individual cell bodies. The calcium trace for each ROI was calculated as the mean fluorescence signal averaged across all pixels comprising the ROI minus the mean fluorescence signal in an annular region surrounding each ROI (neuropil). $\Delta F/F$ was computed as $(F(t) - F_0)/F_0$, where $F(t)$ was the raw calcium snippet surrounding a single tone presentation and F_0 was the mean baseline fluorescence of each snippet during the 0.5 s preceding tone stimulation. For each neuron, we first measured whether $\Delta F/F$ was significantly elevated following tone presentation independently at each frequency ($P < 0.005$, t -test).

Population-level analysis. The calcium traces of all ROIs within an imaging session that were responsive to a particular frequency were averaged together independently for tones presented during running and during rest. Across days, the population responsive to a particular frequency did not necessarily comprise the exact same neurons. To compute the reafferent suppression index (RSI), we first calculated the difference between the tone-evoked calcium traces measured during rest and during running. From the differences measured at the reafferent frequency (α) and a frequency two octaves higher or lower (β), we calculated the RSI as $(\alpha - \beta)/(\alpha + \beta)$. We then tracked RSI as a function of day following the onset of aVR experience.

Single-neuron analysis. Using anatomical coordinates and cell-body morphology, we tracked 577 neurons (from two mice) across subsequent pairs of days. Of these 577 neurons, 241 had significant tone-evoked responses to one or more tones on one or both days. For each neuron, we estimated tuning curves during rest and

during locomotion on each of the two days. We then subtracted the day 1 tuning curve from the day 2 tuning curve, independently for tuning curves measured during rest and locomotion, to determine how much the tuning curve of individual neurons changed across subsequent days of aVR experience. We averaged the change in resting and change in running tuning curves across all 241 tone-responsive neurons.

Single-frequency tone detection task. Mice were acclimated to the treadmill for 2 to 3 days then water restricted for 24 h before training. To train mice to lick in response to tone presentations, tones (25 ms long, 70 and 80 dB SPL) of a fixed frequency (8 kHz) were presented with variable inter-tone-intervals (10 to 15 s) followed by a water reward 1 s later. Licking behaviour was measured with a custom-built infrared detector located between the mouse's mouth and the water delivery spout and was sampled with a data acquisition card (NI) connected to a computer (Dell) running custom software (Matlab). Within 3 to 4 days of training, mice learned to associate the tone pips with rewards. Following this initial training phase, tones were presented at variable inter-tone intervals (8 to 15 s) and mice were required to lick within a 1 s window following tone presentation to receive a reward. To ensure that mice did not lick continuously (to maximize rewards), the inter-tone-interval was reset if the mouse licked during a varying window (5 to 8 s) before tone presentation. Tones of lower intensities (0 to 60 dB, in steps of 10 dB) were gradually introduced over the subsequent 5 days.

During testing, tones were presented using a block design where each block consisted of all tone intensities (0 to 80 dB in steps of 10 dB) presented in random order. To ensure that the mice were engaged in the task, blocks during which the mouse did not lick were removed from analysis. Most mice spontaneously transitioned between periods of running and rest but a subset of mice tended to either rest or run continuously throughout each session. For these mice, tones were only presented during the less explored behaviour, and this transient adjustment of the task criteria was sufficient to alter a mouse's behaviour such that they began to spontaneously transition between periods of running and resting. Mice performed an average of 215 trials per day, consisting of, on average, 170 trials during rest and 45 trials during running. To create psychometric functions for each mouse, trials were separated based on whether mice were running or resting at the time of tone presentation and trials were pooled across all testing days (12.4 ± 5.4 days). 0 dB trials were used to estimate the rate of spontaneous licking and the false alarm rate. For each intensity, hit rate was calculated as the number of correct detections divided by the total number of tone presentations at each intensity. Behavioural threshold was calculated by estimating the intensity at which performance was 50% in each condition. This value was determined by linear interpolation of surrounding intensity values. To account for differences in thresholds due to difference in false alarm rates between conditions, we removed the fraction of correct trials that could be accounted for by false alarms for each condition²⁸. Not applying this correction, and estimating thresholds on our raw data did not change our results (data not shown). Following training each day, mice received 1.5 ml of water. Each mouse was weighed daily to ensure that its weight did not fall below 80% of its pre-water-restriction weight.

Two-frequency tone detection task. Training and testing methods were similar to the single-frequency version of the task, but using tones of two frequencies (4 and 16 kHz). During behavioural testing, trials were presented in a block design in which each block consisted of each frequency-intensity combination (4 and 16 kHz; 0, 30, 40, 50 and 60 dB). Following an initial phase of behavioural testing, the lickometer and water spout were removed from the treadmill and mice received aVR experience with one of the two testing frequencies for 7 days, 2 h per day. Five mice each received aVR experience with 4 kHz and 16 kHz; results were consistent regardless of the aVR frequency used (data not shown). After 7 days of aVR experience, behavioural testing was performed again for 4 to 8 days. At the beginning of each day, mice received 30 min of experience with aVR, during which time they ran and heard the reafferent frequency they were acclimated to over the previous 7 days. This brief re-exposure to aVR was followed by 30 min in their home cage without any aVR or behavioural testing. At the end of this 1-h period, we reintroduced them to the behavioural testing chamber where mice performed the behavioural task.

Optogenetic stimulation during psychophysics. Mice were first trained on the single-tone detection task as described above using a subset of intensities (0, 30, 40 and 60 dB). During behavioural testing, a blue (420 nm) laser (Shanghai) was coupled to a pair of optical fibres (Doric optical splitter), which were positioned bilaterally over the auditory cortex cranial windows. Laser pulses (25 Hz, 50% duty cycle, 15–30 mW) were presented on 50% of sound trials. Laser stimulation began 200 ms before tone presentation and continued for 1.2 s, which covered the entire response window. To accurately estimate the mouse's chance performance on optogenetic trials and to deter the mouse from using light as a cue, 50% of stimuli were laser-only trials, and a short air puff (100 ms, Picospritzer II) was directed towards the mouse's face as negative reinforcement on trials where the animal licked in response to laser-only trials.

Pharmacological manipulation during psychophysics. Mice were first trained on the single-tone detection task as described above. Following initial training, saline or muscimol (2 µg/µl, 150 nl) was bilaterally pressure injected using a Nanoject system into the auditory cortex on alternate days. Mice were allowed to return to their home cage for 30 min after injection. At the end of this period, the mice were reintroduced to the testing chamber where they performed the behavioural task.

Statistical methods. Paired and unpaired two-sided statistical tests were performed with the non-parametric Wilcoxon sign rank test and Wilcoxon rank sum test, respectively, unless otherwise stated. All error bars are s.e. unless otherwise stated. Bootstrap analyses with 1,000 repetitions were used to measure confidence intervals for linear regressions (Fig. 3g, h; Extended Data Fig. 5e–i). Shuffled analyses with 1,000 repetitions were used to estimate null distributions (Fig. 1h, j). Repeated measures two-way ANOVAS followed by post-hoc Tukey tests (where required) were used to compare psychometric curves (Fig. 4c, f, h–j; Extended Data Fig. 6e–g, i–k). ANOVA *P* values were corrected using the Holm–Bonferroni method to account for multiple comparisons (Fig. 4f, h–j; Extended Data Fig. 6j, k). In all figures and captions, *n* is the number of neurons in the data sample and *N* is the number of animals contributing data points to the distributions. No statistical methods were used to predetermine sample sizes, but our sample sizes were similar to those reported in previous publications in the field. Details regarding sample sizes, *P* values and statistical tests for individual figures panels are detailed below. **Figure 1f.** Of 317 neurons (*n* = 11 mice), 120 were significantly responsive to the reafferent frequency and 248 were responsive to at least one non-reafferent frequency, calculated by comparing baseline firing rates during the 100 ms preceding tone onset to the 100 ms following tone onset ($P < 0.005$, paired *t*-test). During rest, response strength (driven minus baseline firing rates) to reafferent and non-reafferent tones were not significantly different (11.2 ± 22.3 vs. 13.1 ± 24.1 action potentials/s, *n* = 120 and 248, $P = 0.46$). During running, response strength to the reafferent tone was significantly less than the response strength to non-reafferent tones (-0.1 ± 19.9 vs. 6.3 ± 23.3 action potentials/s, *n* = 120 and 248, $P = 0.001$). For a paired, within-neuron comparison, we restricted our analysis to neurons that responded to the reafferent frequency and at least one other frequency (*n* = 115) and we averaged each neuron's response to all non-reafferent frequencies to which it responded significantly. During running, responses to the reafferent frequency were weaker than to non-reafferent frequencies ($P = 1.1 \times 10^{-18}$, Wilcoxon signed-rank test). As a population, responses to the reafferent frequency during running were not significantly different from 0 ($P = 0.12$).

Figure 1h. We aligned the gain function for each neuron to the reafferent frequency experienced by the mouse from which the neuron was recorded. The number of neurons responsive to each tone frequency (−3, −2, −1, 0, 1, 2, and 3 octaves, relative to the reafferent frequency) were 70, 92, 153, 120, 110, 95, and 67, respectively. To determine whether the notch at the reafferent frequency was significant relative to what would be expected by chance, we shuffled the data by randomly assigning to each neuron a reafferent frequency rather than using the actual frequency experienced by the mouse from which the neuron was recorded. This shuffling was performed 1,000 times and the 95% confidence bounds of the distribution were computed.

Figure 1j. As in Fig. 1h. 109 neurons were recorded from five mice. The number of neurons responsive to each tone frequency (−2, −1, 0, 1, and 2 octaves, relative to the reafferent frequency) were 24, 39, 33, 24, and 20, respectively. As in Fig. 1h, we computed confidence bounds by randomly assigning to each neuron a reafferent frequency rather than using the actual frequency experienced by the mouse from which the neuron was recorded.

Figure 2c. Solid lines show the mean calcium response to tones averaged across all neurons that were responsive to that tone during rest. The population size that was responsive to each tone on each day is noted below each set of red and black traces. Vertical line shows dF/F .

Figure 2i. For 241 neurons, the tuning curves estimated across subsequent days were subtracted, and this was done independently for curves measured during rest (black) and running (red). At each frequency (and independently for each movement condition), we performed a bootstrap analysis, randomly sampling from our distribution (with replacement), repeated 1,000 times. The shaded areas show the 95% confidence bounds for these distributions. The only significant change in tuning was at 32 kHz (the reafferent frequency) and only during locomotion (red).

Figure 3d. Data are from pi-INs recorded from two *PV::Cre* mice (*n* = 36), two *SST::Cre* mice (*n* = 21), and three *VGAT::ChR2* mice (*n* = 44). Magenta, green and black asterisks indicate $P = 0.03$, $P = 0.04$, and $P = 0.02$, respectively. Analysing all interneurons together, $P = 0.01$. (sign rank and rank sum tests for paired and unpaired tests, respectively).

Figure 3g. Data are from pi-INs recorded from two *PV::Cre* mice (*n* = 54), two *SST::Cre* mice (*n* = 32), and five *VGAT::ChR2* mice (*n* = 75). Solid lines are linear regression and shaded lines show 95% confidence bounds from a bootstrap analysis repeated 1,000 times.

Figure 3h. Data are from pi-INs recorded from two *PV::Cre* mice (*n* = 33) and two *SST::Cre* mice (*n* = 26). PV^+ and SST^+ neurons were pooled for the regression analysis. Solid lines are linear regression and shaded lines show 95% confidence bounds from a bootstrap analysis repeated 1,000 times.

Figure 3i. Data are from regressions shown in g and Extended Data Fig. 5g–i. Effect sizes for PV, VGAT, SST, and all pi-INs are significantly larger than effect sizes for non-reafferent and naive conditions ($P < 0.01$, Wilcoxon). Effect sizes for PV, VGAT and all pi-INs are significantly larger than effect size for put-ENs ($P < 0.01$, Wilcoxon). Bar height determined by linear fit of raw data; error bars show s.e. of linear fits from 1,000 repetitions of bootstrap analysis.

Figure 4c. Resting and running performance curves are mean and s.e. for *N* = 19 mice. Lines are sigmoid fits to the data points. Comparisons with repeated measures two-way ANOVAS at non-zero intensities (factors: intensity \times behavioural state, $P(\text{interaction}) = 0.0003$, $F(7, 126) = 4.22$), followed by post-hoc Tukey test. Red asterisks, $P < 0.005$.

Figure 4d. Each point contains the behavioural threshold (intensity at 50% performance) during rest and running for a single mouse. Comparisons with two-sided paired *t*-test. $P = 0.009$.

Figure 4f. Data points show mean and s.e. for *N* = 4 mice. Rest versus optogenetic activation of inhibitory interneurons in the auditory cortex: comparisons with repeated measures two-way ANOVAS at non-zero intensities. (factors: intensity \times laser state, $P(\text{interaction}) = 0.01$, $F(2, 6) = 14.27$, post-hoc Tukey test, blue asterisk $P < 0.05$). Rest versus running: comparisons with repeated measures two-way ANOVAS at non-zero intensities (factors: intensity \times behavioural state, $P(\text{interaction}) = 0.03$, $F(2, 6) = 6.60$, post-hoc Tukey test, red asterisk $P < 0.05$). ANOVA *P* values were corrected using the Holm–Bonferroni method.

Figure 4h. Data points show mean and s.e. for *N* = 4 mice. Rest versus optogenetic activation of M2 axon terminals in the auditory cortex: comparisons with repeated measures two-way ANOVAS at non-zero intensities (factors: intensity \times laser state, $P(\text{interaction}) = 0.04$, $F(2, 6) = 5.84$, post-hoc Tukey test, blue asterisk, $P < 0.05$). Rest versus running: comparisons with repeated measures two-way ANOVAS at non-zero intensities. (factors: intensity \times behavioural state, $P(\text{interaction}) = 0.04$, $F(2, 6) = 8.29$, post-hoc Tukey test, red asterisk, $P < 0.05$). ANOVA *P* values were corrected using the Holm–Bonferroni method.

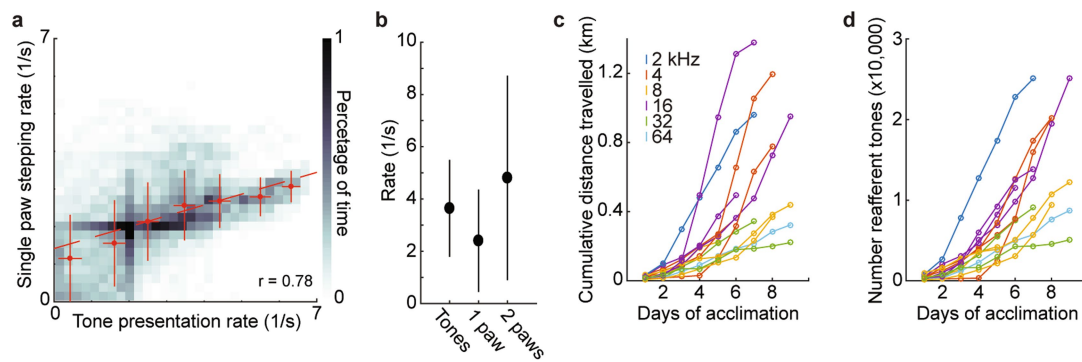
Figure 4i. Data points show mean and s.e. for *N* = 10 mice during rest and running. Comparisons in each subpanel with repeated measures two-way ANOVAS at non-zero intensities (factors: intensity \times behavioural state. For tone A: $P(\text{interaction}) = 0.0002$, $F(3, 27) = 11.56$. For tone B: $P(\text{interaction}) = 0.04$, $F(3, 27) = 3.85$), followed by post-hoc Tukey test. ANOVA *P* values corrected using the Holm–Bonferroni method. Red asterisks, $P < 0.05$.

Figure 4j. Data points show mean and s.e. for *N* = 10 mice during rest and running. Comparisons with repeated measures two-way ANOVAS at non-zero intensities (factors: intensity \times behavioural state. For tone A: $P(\text{interaction}) = 0.01$, $F(3, 27) = 5.62$. For tone B: $P(\text{interaction}) = 0.05$, $F(3, 27) = 2.90$), followed by post-hoc Tukey test. ANOVA *P* values were corrected using the Holm–Bonferroni method. Red asterisks, $P < 0.05$.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

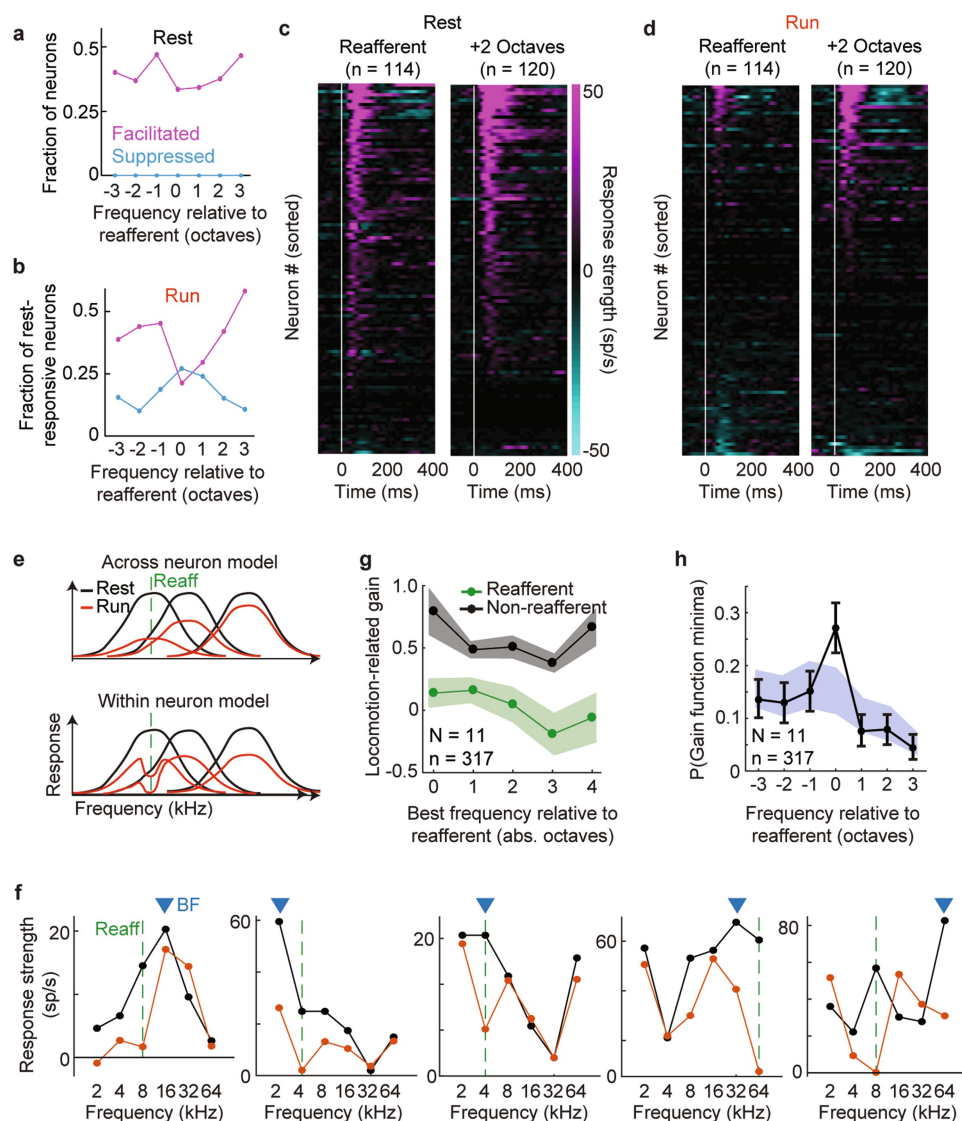
28. Glickfeld, L. L., Histed, M. H. & Maunsell, J. H. Mouse primary visual cortex is used to detect both orientation and contrast changes. *J. Neurosci.* **33**, 19416–19422 (2013).



Extended Data Fig. 1 | aVR experience is coupled to locomotion.

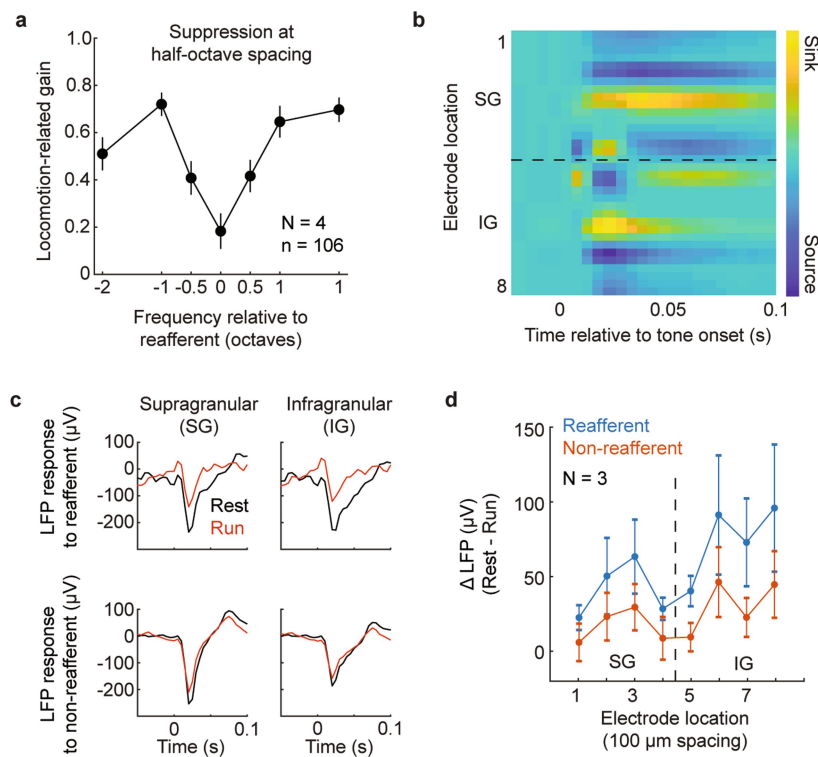
a, Heat map showing the rate of tone presentation as a function of instantaneous stepping rate with a single paw, measured via simultaneous videography. Data points show mean \pm s.d. tone rate and stepping rate in 1-Hz (1 s^{-1}) bins. Red dashed line shows linear regression through all data points. Refferent tones during aVR experience were strongly correlated to instantaneous paw stepping rate (0.78). Data are from 3,716

steps recorded from 1,804 s of video from two mice. **b**, Average tone presentation rate during aVR experience closely matches average stepping rate measured either with a single paw or two paws. Dots are median and error bars are s.d. **c**, Cumulative distance run by 11 mice over 6–9 days of aVR experience. Each line is for a different mouse, colour-coded by the refferent frequency to which the mouse was acclimated. **d**, Cumulative number of tones heard by same 11 mice as in c.



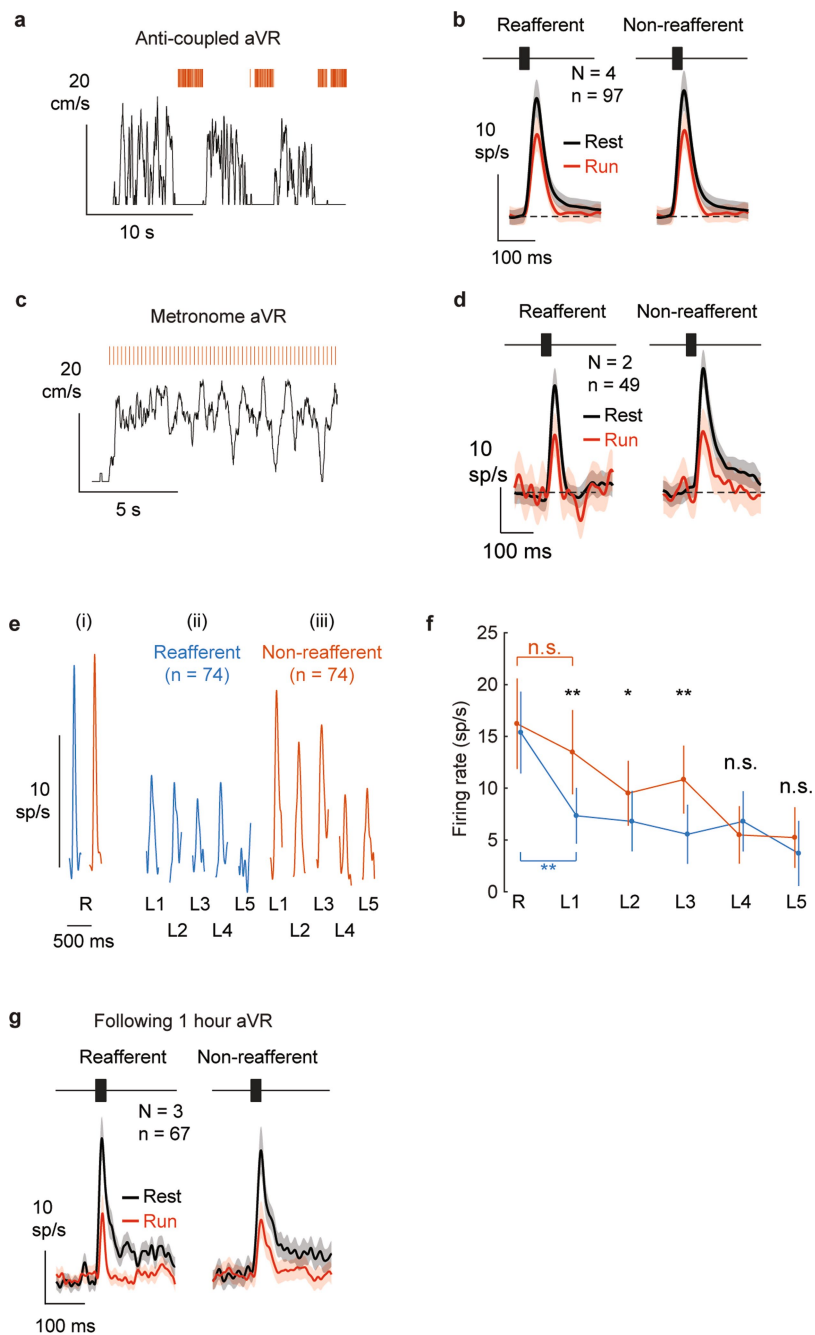
Extended Data Fig. 2 | aVR experience alters locomotion-related suppression at the level of individual neurons. **a**, Fraction of neurons with elevated firing rates (magenta) and suppressed firing rates (cyan) in response to tones during rest. A roughly equal number of neurons were excited by the reafferent frequency as were excited by other frequencies. **b**, Fraction of rest-responsive neurons with elevated firing rates (magenta) and suppressed firing rates (cyan) in response to tones of varying frequency during running. Nearly 50% of neurons were responsive to non-reafferent frequencies during running, whereas fewer than 25% were responsive to the reafferent frequency. **c**, Heat map showing response strength (tone-evoked rate – baseline rate) for neurons responsive to the expected reafferent frequency (left, $n = 114$ neurons, $N = 11$ mice) and another frequency (+2 octaves, $n = 120$ neurons, $N = 11$ mice) during rest. Neurons ordered by magnitude of response independently for each heat map. **d**, Response strengths of the neurons in **c** during running. Neurons are re-sorted by magnitude of response. Twenty-three per cent of neurons retained their response to the reafferent frequency during running, consistent with a sparse representation of expected reafferent sounds. **e**, Two alternative models for how locomotion-related suppression could change following aVR experience. In each model, the black curves show frequency tuning curves of three neurons during rest, red curves during running, and the green dashed line indicates the reafferent frequency. Across-neuron model: locomotion-related suppression is uniform across frequencies within a neuron but is strongest for neurons that are strongly responsive to the expected reafferent frequency. Within-neuron model: suppression is non-uniform at the single neuron level and regardless of how strongly the neuron responds to the expected reafferent frequency,

suppression is always strongest at the reafferent frequency. **f**, Tuning curves for five example neurons measured during rest (black) and running (red). The best frequency (BF) for each neuron is shown by the blue triangle, and the reafferent frequency to which each mouse was acclimated is shown by the green dashed line. In all five neurons, locomotion-related suppression was strong at the reafferent frequency relative to other frequencies, regardless of the neuron's best frequency. **g**, Neurons were sorted by their best frequency, measured relative to the reafferent frequency that each mouse experienced. Locomotion-related suppression at the expected reafferent frequency (green) and averaged across all non-reafferent frequencies (black). Regardless of a neuron's best frequency, suppression was always strongest at the reafferent frequency, supporting the within-neuron model in **e**. Sample size: $N = 11$ mice, $n = 314$ neurons. Shaded regions show 95% confidence bounds estimated with a bootstrap analysis repeated 1,000 times. **h**, Probability of observing a minima in the gain function of individual neurons at each frequency, measured relative to the reafferent frequency. A substantial number of neurons had minima in their gain functions at the expected reafferent frequency, further supporting the within-neuron model in **e**. Sample size: $N = 11$ mice, $n = 314$ neurons. Shaded region shows a null distribution, which we estimated by randomly assigning to each neuron a reafferent frequency rather than using the actual frequency experienced by the mouse from which the neuron was recorded. This shuffling was performed 1,000 times and the 95% confidence bounds of the distribution were computed. Error bars show the 95% confidence bounds estimated from a bootstrap analysis repeated 1,000 times.



Extended Data Fig. 3 | Specificity of suppression following aVR experience. **a**, Locomotion-related gain tested at half-octave spacing from the reafferent frequency. Neuronal responses to frequencies half an octave from the reafferent frequency were suppressed at an intermediate level. Data are mean \pm s.e. Sample size: $N = 4$ mice, $n = 106$ neurons. **b**, Example current-source density triggered by tone-onset for electrode recordings made perpendicular to the auditory cortical surface. Black dashed line demarcates putative supragranular (SG) and infragranular (IG) layers of cortex. Electrode 1 is the most superficial; electrode spacing

is 100 μm . **c**, Example tone-evoked local field potential (LFP) traces from an SG electrode (left) and an IG electrode (right) in response to the expected reafferent frequency (left) and a non-reafferent frequency (right). Locomotion-related suppression of LFP responses was stronger for the reafferent frequency than for non-reafferent frequencies. Data are mean \pm s.e. **d**, The difference in LFP between rest and running as a function of electrode location (1 is the most superficial; electrode spacing is 100 μm ; $N = 3$ mice). Positive values indicate greater suppression during locomotion.

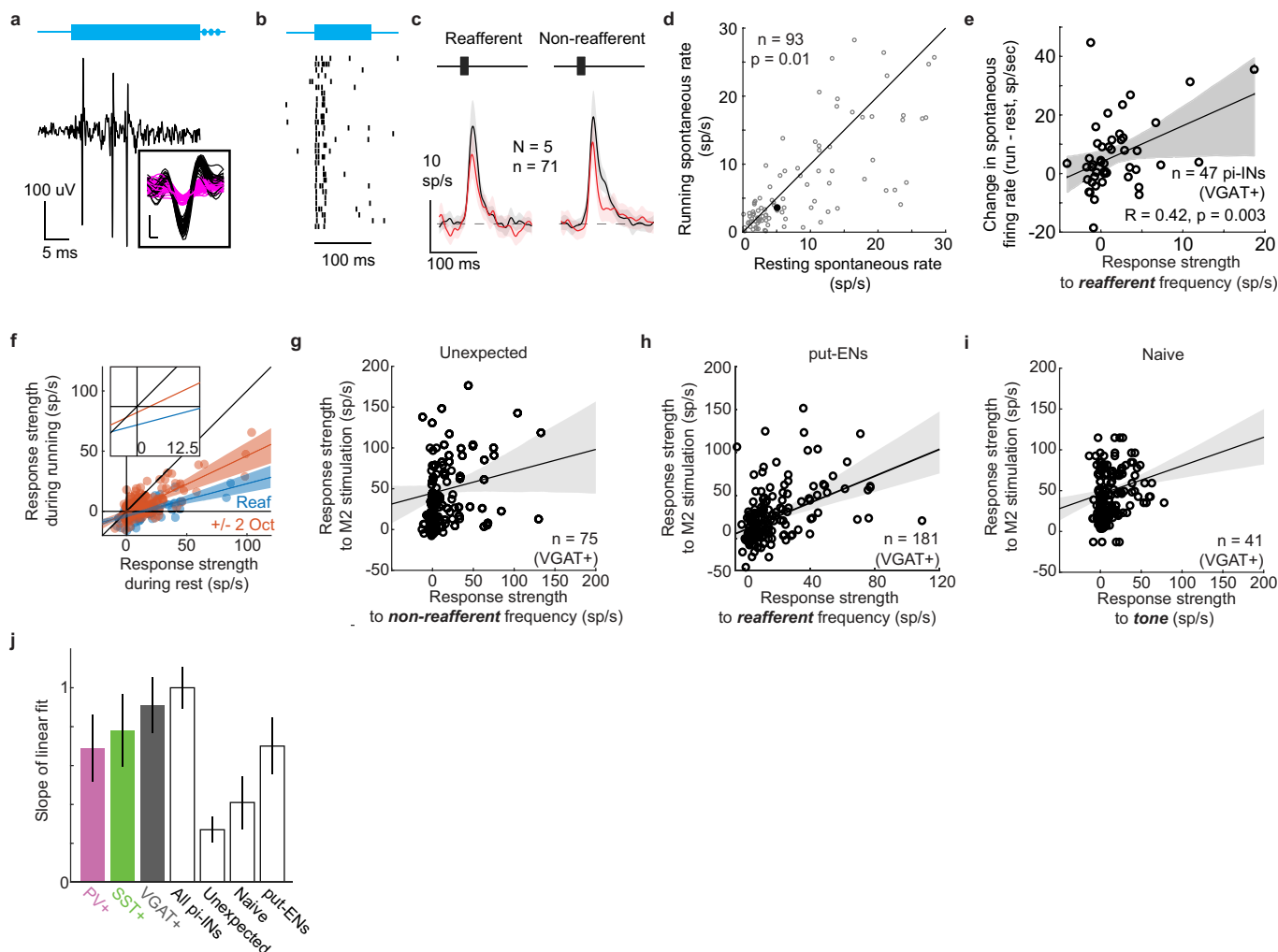


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Frequency-specific locomotion-related suppression requires several days of coupled sensory-motor experience.

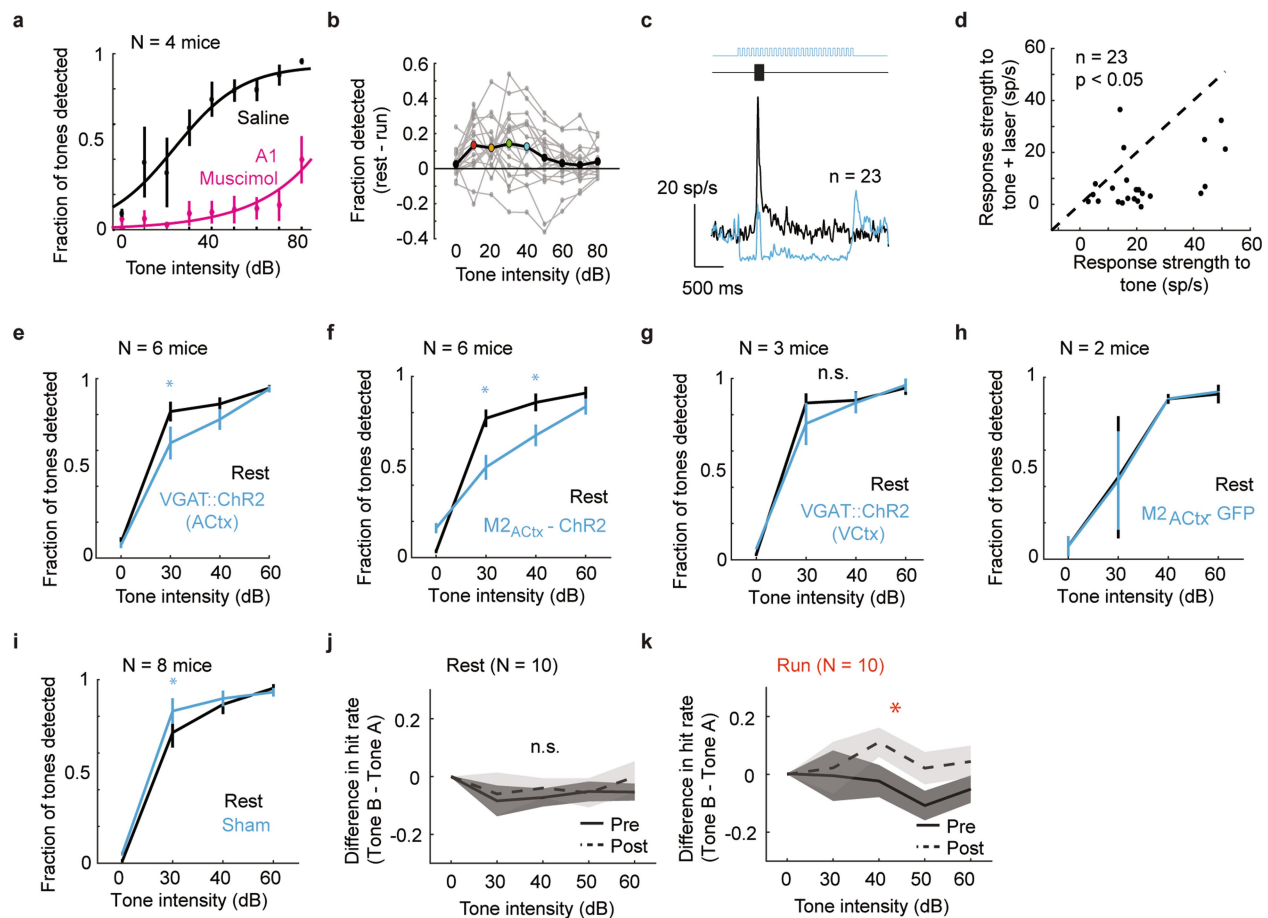
a, Example sensory-motor experience during anti-coupled aVR experience. Mice did not hear tones while running, but tones were played back during subsequent resting periods with inter-tone intervals drawn from the intervals that mice should have heard while running. **b**, Population PSTHs for the expected frequency (left) and for non-reafferent frequencies (right) during rest (black) and running (red) following anti-coupled aVR. Anti-coupled aVR experience does not lead to changes in auditory responsiveness during running or rest. Sample size: $N = 4$ mice, $n = 97$ neurons. Shaded region shows mean \pm s.e. $P = 0.57$, two-sided Wilcoxon rank sum test. **c**, Example sensory-motor experience during metronome aVR experience. Tones were presented during running at a fixed rate (2 s^{-1}) but the tone rate was not modulated by running speed. **d**, Population PSTHs for the expected frequency (left) and for non-reafferent frequencies (right) during rest (black) and running (red) following metronome aVR. Metronome aVR experience does not lead to changes in auditory responsiveness during running or rest. Sample size: $N = 2$ mice, $n = 49$ neurons. Shaded region shows mean \pm s.e. $P = 0.57$, two-sided Wilcoxon rank sum test. **e**, Mice were acclimated to aVR for 7 days. On the day of electrophysiology, we altered on each locomotor bout the sound produced by the treadmill to be either expected (blue) or a non-reafferent frequency (2 octaves away, red). We then analysed responses ($N = 4$ mice, $n = 74$ neurons) to each sound frequency during rest (R) and to the first five tones heard at the beginning of each bout of

locomotion (L1–L5). (i) Tone-evoked responses (population PSTHs) to the reafferent (blue) and a non-reafferent sound (red) during rest. (ii) Tone-evoked responses during locomotion to the first five tones in a series of the expected reafferent frequency. (iii) Tone-evoked responses during locomotion to the first five tones heard in a series of non-reafferent tones. **f**, Firing rates to the reafferent (blue) and non-reafferent (red) reafferent sounds during rest (R) and during the first five tones heard during locomotion (L1–L5). Responses to the first tone heard during locomotion were significantly suppressed only if that tone matched the expected reafferent frequency (blue asterisk, $P = 0.002$, two-sided Wilcoxon signed rank test). Black asterisks indicate significant differences between firing rates to the reafferent and non-reafferent reafferent sounds (L1, $P = 0.002$; L2, $P = 0.03$; L3, $P = 0.007$, two-sided Wilcoxon rank sum test). Sample size: $N = 4$ mice, $n = 74$ neurons. Red n.s. indicates that evoked responses to the first tone heard during a bout of running are not significantly different from those evoked during rest for non-reafferent tones ($P = 0.4$, two-sided Wilcoxon signed rank test). **g**, Population PSTHs for the expected frequency (left) and for non-reafferent frequencies (right) during rest (black) and running (red). Data were collected from three mice ($n = 67$ neurons) after each mouse's first experience of hearing fixed-frequency reafferent tones for 1 h, during which time mice heard 927, 3,167 and 1,069 reafferent tones at 16 kHz, 2 kHz and 16 kHz, respectively. This experience was insufficient to shift the locomotion-related suppression towards the reafferent frequency. Shaded region shows mean \pm s.e. $P = 0.47$, two-sided Wilcoxon rank sum test.



Extended Data Fig. 5 | Characterizing photo-identified inhibitory neurons in auditory cortex. **a**, Voltage trace of a pi-IN recorded from a *VGAT::ChR2* mouse in response to a 100-ms pulse of blue light targeted to the cortical surface. Inset shows example waveforms belonging to the sorted unit (black) and belonging to the noise cluster (magenta), showing good electrophysiological isolation. **b**, Rasters showing response of the same neuron to 30 pulses of blue light (100 ms each). **c**, Tone-evoked responses of auditory cortical inhibitory neurons (*VGAT*⁺) during rest (black) and locomotion (red) in response to refferent (left) and non-refferent (right) frequencies. Responses are suppressed during locomotion, but suppression is not specific to the refferent frequency. Sample size: $N = 5$ mice, $n = 71$ neurons. Shaded region shows mean \pm s.e. $P = 0.36$, two-sided Wilcoxon rank sum test. **d**, Spontaneous firing rate during rest and locomotion for 93 putative excitatory neurons (non-photo-identified in *VGAT::ChR2* mice, $N = 7$ mice). Filled circle shows mean. Firing rates were significantly lower during running relative to rest (two-sided Wilcoxon signed rank test). **e**, pi-INs (*VGAT*⁺) that were more strongly driven by the refferent frequency were more strongly recruited during running. $N = 2$ mice, $n = 47$ neurons. Black line and shaded area show linear regression and 95% confidence bounds from a bootstrap analysis repeated 1,000 times, respectively. The P value represents the probability that the slope of the regression line includes zero, estimated from the bootstrap analysis. **f**, Tone-evoked responses during running and rest for the refferent frequency (blue) and non-refferent frequencies (± 2 octaves, red). Dots are responses of individual neurons ($N = 11$ mice, $n = 317$), lines are linear regression, and shaded regions are 95% confidence bounds from bootstrap analysis repeated 1,000 times. Suppression to non-refferent sounds is best fit as a gain model (slope = 0.47 ± 0.05 ; offset = -0.19 ± 0.70), whereas suppression of expected refferent tones has a stronger gain component (that is,

shallower slope, two-sided Wilcoxon rank sum test, $P = 3.3 \times 10^{-317}$) and an offset term that is significantly different from zero (slope = 0.27 ± 0.4 ; offset = -3.55 ± 0.58 , two-sided signed rank test, $P = 3.3 \times 10^{-163}$). Inset shows a zoom in of the regression lines near the origin. These data suggest that suppression of expected refferent sounds involves both divisive and subtractive forms of inhibition. **g**, Responses to a non-refferent tone in *VGAT*⁺ pi-INs recorded from aVR-acclimated mice were weakly correlated with responses to electrical stimulation in M2 ($n = 75$ neurons from 5 mice). These data indicate that the strong relationship between tone-evoked responses and M2 stimulation responses in auditory cortical pi-INs is distinct to the refferent frequency. Black line and shaded area show linear regression and 95% confidence bounds from a bootstrap analysis repeated 1,000 times, respectively. **h**, Responses to the expected refferent tone in put-ENs recorded from aVR-acclimated mice were correlated with responses to electrical stimulation in M2 ($n = 181$ neurons from 5 *VGAT::ChR2* mice). This effect size for put-ENs is significantly weaker than for pi-INs. Black line and shaded area show linear regression and 95% confidence bounds from a bootstrap analysis repeated 1,000 times, respectively. **i**, Responses to a non-refferent tone in *VGAT*⁺ pi-INs recorded from naive mice were weakly correlated with responses to electrical stimulation in M2 ($n = 41$ neurons from 2 mice). Black line and shaded area show linear regression and 95% confidence bounds from a bootstrap analysis repeated 1,000 times, respectively. **j**, Slope of the linear fit for the relationship shown in Fig. 3i. Error bars show 95% confidence bounds from a bootstrap analysis. Data are from regressions shown in Fig. 3g and Extended Data Fig. 5g–i. Slopes of linear fit for PV, *VGAT*, SST, and all pi-INs are significantly larger than slopes of linear fits for non-refferent and naive conditions ($P < 0.01$, Wilcoxon). Bar height determined by linear fit of raw data; error bars show s.e. of linear fits from 1,000 repetitions of bootstrap analysis.



Extended Data Fig. 6 | Tone detection behaviour is compromised by locomotion, is auditory-cortex dependent, and adapts following VR experience. **a**, Data points show mean and s.e. detection rates for $N = 4$ mice as a function of tone intensity for trials performed during rest with infusion of either saline (black) or muscimol (magenta) into the auditory cortex. **b**, Difference in performance as a function of intensity for each mouse (grey dots). Large connected dots show mean difference in performance and coloured dots indicate intensities at which performance was significantly different ($P < 0.05$) across conditions ($N = 19$ mice, repeated measures two-way ANOVA followed by post-hoc Tukey test). **c**, Tone-evoked responses from putative excitatory neurons recorded from VGAT::ChR2 mouse without (black) and with (blue) simultaneous blue laser stimulation. Optogenetic activation of inhibitory neurons decreases the spontaneous and tone-evoked firing rates of excitatory neurons. $n = 23$ neurons, $N = 1$ mouse. **d**, Tone-evoked firing rates of inhibitory interneurons. Dashed line is unity. ($n = 23$ neurons, $N = 1$ mouse; $P < 0.05$, two-sided paired t -test). **e**, Tone detection performance ($N = 6$ mice) during rest (black) and rest with optogenetic activation of auditory cortical inhibitory neurons (blue). Mice were worse at detecting tones on optogenetic trials (repeated measures two-way ANOVA, factors: intensity \times laser state, $P(\text{intensity} \times \text{laser state}) = 0.0028$, $F(2, 10) = 11.23$, post-hoc Tukey test at individual intensities, blue asterisk, $P < 0.05$ on laser trials) compared to rest. **f**, Tone detection performance ($N = 6$ mice) during rest (black) and rest with optogenetic activation of M2 terminals in auditory cortex (blue). Four of these mice were presented with 8-kHz tones and the remaining two were presented with 4-kHz tones. Mice were worse at detecting tones on optogenetic trials regardless of the tone frequency. (Statistics similar to **e**, $P(\text{intensity} \times \text{laser state}) = 0.01$, $F(2, 10) = 6.66$, blue asterisk, $P < 0.05$ on laser trials). **g**, Average psychometric functions ($N = 3$ mice) showing detection rates as a function of tone intensity for trials performed during rest when visual cortex was inhibited. (repeated measures

two-way ANOVA, $P(\text{intensity} \times \text{laser state}) = 0.33$, $F(2, 4) = 1.47$). **h**, Average psychometric functions ($N = 2$ mice) showing detection rates as a function of tone intensity for trials performed during rest (black) and during rest with laser stimulation (blue) by mice injected with an AAV encoding eGFP in M2. These controls show that laser stimulation of auditory cortex in the absence of ChR2 does not influence behaviour. **i**, Average psychometric functions ($N = 8$ mice) showing detection rates as a function of tone intensity for trials performed during rest (black) and during rest with laser stimulation (blue) when the optical fibre was placed over intact skull near, but not directly over auditory cortex. Five of eight mice were injected with an AAV encoding ChR2 into M2, of which three were presented with 8-kHz tones and 2 with 4-kHz tones. The other three were VGAT::ChR2 mice presented with 8-kHz tones. These controls show that sham laser stimulation (which is visible to the mouse) alone improves behaviour (repeated measures two-way ANOVA, factors: intensity \times laser state, $P(\text{interaction}) = 0.0066$, $F(2, 14) = 7.35$, post-hoc Tukey tests, blue asterisk, $P < 0.05$). **j**, Difference in hit rates in response to tone A relative to tone B during rest before (pre) and after (post) aVR experience with tone A. Lines represent mean difference and shaded regions show s.e. for $N = 10$ mice. There is no difference in rest performance before and after aVR experience. (repeated measures two-way ANOVA in each panel, factors: intensity \times time of testing, $P(\text{time of testing}) = 0.46$, $F(1, 9) = 0.61$). **k**, Difference in hit rates in response to tone A relative to tone B during running before (pre) and after (post) aVR experience with tone A. Lines represent mean difference and shaded regions show s.e. for $N = 10$ mice. Mice are significantly better at detecting tone B than tone A after aVR experience, indicating that this is a movement-specific change (repeated measures two-way ANOVA in each panel, factors: intensity \times time of testing, $P(\text{time of testing}) = 0.04$, $F(1, 9) = 8.07$, red asterisk, $P < 0.05$, p values in **j**, **k** corrected using the Holm-Bonferroni method. For further statistical details, see Supplementary Table 1.

Required growth facilitators propel axon regeneration across complete spinal cord injury

Mark A. Anderson^{1,2,6}, Timothy M. O'Shea^{1,6}, Joshua E. Burda¹, Yan Ao¹, Sabry L. Barlatey², Alexander M. Bernstein¹, Jae H. Kim¹, Nicholas D. James², Alexandra Rogers¹, Brian Kato¹, Alexander L. Wollenberg³, Riki Kawaguchi⁴, Giovanni Coppola⁴, Chen Wang⁵, Timothy J. Deming³, Zhigang He⁵, Gregoire Courtine^{2,7*} & Michael V. Sofroniew^{1,7*}

Transected axons fail to regrow across anatomically complete spinal cord injuries (SCI) in adults. Diverse molecules can partially facilitate or attenuate axon growth during development or after injury^{1–3}, but efficient reversal of this regrowth failure remains elusive⁴. Here we show that three factors that are essential for axon growth during development but are attenuated or lacking in adults—(i) neuron intrinsic growth capacity^{2,5–9}, (ii) growth-supportive substrate^{10,11} and (iii) chemoattraction^{12,13}—are all individually required and, in combination, are sufficient to stimulate robust axon regrowth across anatomically complete SCI lesions in adult rodents. We reactivated the growth capacity of mature descending propriospinal neurons with osteopontin, insulin-like growth factor 1 and ciliary-derived neurotrophic factor before SCI^{14,15}; induced growth-supportive substrates with fibroblast growth factor 2 and epidermal growth factor; and chemoattracted propriospinal axons with glial-derived neurotrophic factor^{16,17} delivered via spatially and temporally controlled release from biomaterial depots^{18,19}, placed sequentially after SCI. We show in both mice and rats that providing these three mechanisms in combination, but not individually, stimulated robust propriospinal axon regrowth through astrocyte scar borders and across lesion cores of non-neural tissue that was over 100-fold greater than controls. Stimulated, supported and chemoattracted propriospinal axons regrew a full spinal segment beyond lesion centres, passed well into spared neural tissue, formed terminal-like contacts exhibiting synaptic markers and conveyed a significant return of electrophysiological conduction capacity across lesions. Thus, overcoming the failure of axon regrowth across anatomically complete SCI lesions after maturity required the combined sequential reinstatement of several developmentally essential mechanisms that facilitate axon growth. These findings identify a mechanism-based biological repair strategy for complete SCI lesions that could be suitable to use with rehabilitation models designed to augment the functional recovery of remodelling circuits.

We tested the hypothesis that the failure of adult central nervous system (CNS) axons to regrow across complete SCI lesions is due to a combined lack of several mechanisms that are required for developmental axon growth. We targeted descending propriospinal neurons because after incomplete SCI they spontaneously form new intraspinal circuits that relay functionally meaningful information past lesions^{20–22}. Thus, short-distance regrowth of transected propriospinal axons across complete SCI lesions has the potential to find new neuronal targets and form new relay circuits. To reactivate intrinsic propriospinal neuronal growth capacity, which is attenuated in adult CNS neurons^{2,5–9}, we tested two approaches that have previously been successful with retinal and corticospinal neurons: using adeno-associated viral vectors (AAV) to deliver either phosphatase and tensin homologue (PTEN) knockdown (AAV-shPT)²³, or to express

osteopontin, insulin-like growth factor 1 (IGF1) and ciliary-derived neurotrophic factor (CNTF) (AAV-OIC)^{14,15}. To increase axon growth-supportive substrates such as laminin^{10,11,19,24}, we delivered fibroblast growth factor 2 (FGF2)²⁵ and epidermal growth factor (EGF)²⁶. To chemoattract^{12,13} propriospinal axons, we delivered glial-derived growth factor (GDNF) because propriospinal neurons express GDNF receptors (GDNFR), increase GDNFR expression after SCI¹⁶ and regrow axons into GDNF-secreting grafts¹⁷, and because SCI lesions lack GDNF¹⁹. To provide temporally controlled and spatially targeted delivery of growth factors or function-blocking antibodies, we used biomaterial depots of synthetic hydrogels^{18,19} placed sequentially into SCI lesion centres and into caudal spared neural tissue (Fig. 1a, Extended Data Fig. 1). AAV injected one segment rostral to SCI lesions efficiently targeted propriospinal neurons, including neurons expressing GDNFR (Extended Data Fig. 2a, b). These manipulations were tested alone and in combinations, first in adult mice and then in adult rats with severe crush SCI causing anatomically complete lesions, across which there is no spontaneous regrowth of descending or ascending axons¹⁹.

Propriospinal axon regeneration was quantified as tract-tracer-labelled axons that regrew to lesion centres or beyond (Fig. 1b–d, Extended Data Figs. 2c–e, 3, 4). Mice with SCI only or SCI plus empty hydrogel depots exhibited few or no axons at lesion centres. Individual interventions, AAV-shPT or AAV-OIC alone, or depots with FGF and EGF alone or GDNF alone, did not significantly increase this number. Combined delivery of all three growth factors, FGF, EGF and GDNF in one or two depots, but without AAVs, led to modest but significantly increased axon regrowth. Combined delivery of AAV-shPT plus FGF, EGF and GDNF did not significantly increase axon numbers compared with FGF, EGF and GDNF alone or with control AAV delivering a nonsense sequence plus FGF, EGF and GDNF. By contrast, combined delivery of AAV-OIC plus FGF, EGF and GDNF synergistically facilitated robust propriospinal axon regrowth. In mice with two sequentially placed depots, this regrowth passed through non-neural lesion cores and their astrocyte scar borders, and penetrated well into spared grey matter (Figs. 1–3, Extended Data Figs. 3, 4). In these mice, total axon regrowth past lesion centres was over 100-fold greater than SCI only or SCI plus empty depots (Fig. 1c). Regrowing propriospinal axons expressed detectable GDNFR (Fig. 1e) and followed irregular paths (Fig. 1d) consistent with regrowing, as opposed to spared, axons²⁷. Biotinylated dextran amine (BDA) tract-tracer did not label axons of passage, such as serotonin axons (Extended Data Fig. 2d, e). In all cases, no BDA-labelled axons were present 3 mm past lesion centres, confirming that lesions were complete (Fig. 1b–d, Extended Data Figs. 3, 4).

To dissect potential cellular and molecular mechanisms underlying this robust axon regrowth, we first examined axon–substrate interactions (Fig. 2, Extended Data Figs. 5, 6). FGF with EGF significantly

¹Department of Neurobiology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. ²Center for Neuroprosthetics and Brain Mind Institute, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. ³Departments of Bioengineering, Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA, USA. ⁴Departments of Psychiatry and Neurology, University of California, Los Angeles, Los Angeles, CA, USA. ⁵F.M. Kirby Neurobiology Center, Department of Neurology, Children's Hospital, Harvard Medical School, Boston, MA, USA. ⁶These authors contributed equally: Mark A. Anderson, Timothy M. O'Shea. ⁷These authors jointly supervised this work: Gregoire Courtine, Michael V. Sofroniew. *e-mail: gregoire.courtine@epfl.ch; sofroniew@mednet.ucla.edu

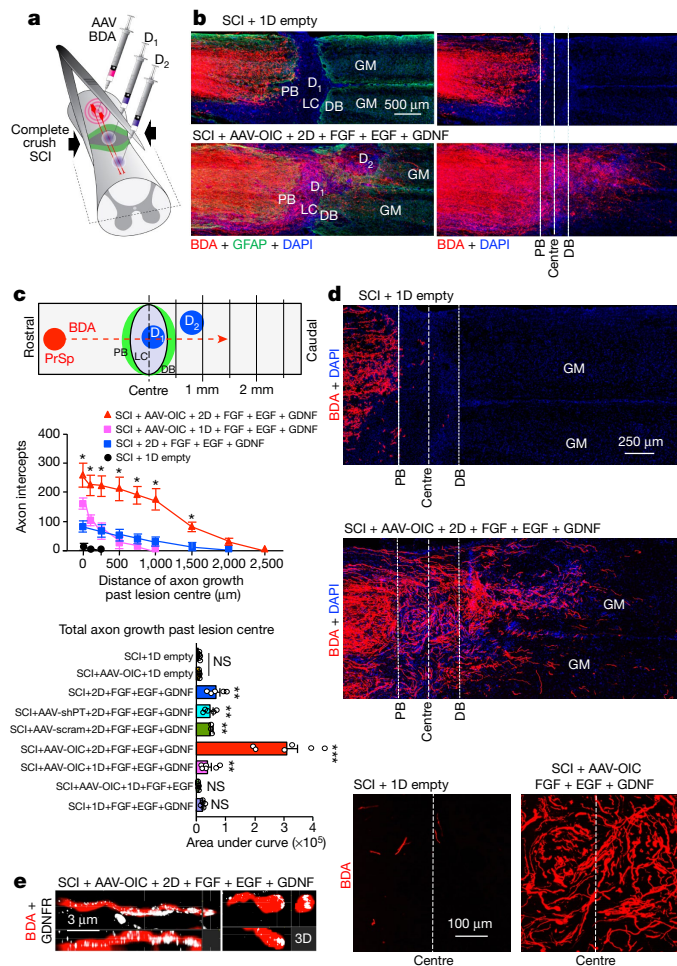


Fig. 1 | Stimulated and chemoattracted propriospinal axons regrow robustly across anatomically complete SCI lesions in mice receiving combined delivery of AAV-OIC plus FGF, EGF and GDNF in two sequentially placed hydrogel depots. **a**, Experimental model. D₁ and D₂ hydrogel depot 1 and 2, respectively. **b**, BDA-labelled axons in composite tiled scans of horizontal sections also stained for astrocytes (anti-GFAP (glial fibrillary acidic protein), left) and cell nuclei (DAPI). Dotted lines demarcate astrocyte proximal (PB) and distal (DB) border around the lesion core (LC). Dashed line demarcates lesion centre. 1D, one depot; 2D, two depots; GM, grey matter. **c**, Top, schematic of axon intercept. Middle, axon intercepts at specific distances past lesion centres (colour coding and *n* as in the graph below). Bottom, areas under axon intercept curves. Dots show *n* mice per group. NS, not significant versus SCI + 1D empty; **P* < 0.01 versus SCI + 1D empty, two-way ANOVA with Bonferroni; ***P* < 0.01 versus all other groups and not significant versus each other, ****P* < 0.0001 versus all other groups, one-way ANOVA with Bonferroni. Data are mean ± s.e.m. PrSp, propriospinal. **d**, Surveys (top) and details (bottom) of BDA-labelled axons. **e**, Three-dimensional detail of BDA-labelled axon and growth cone expressing GDNFR in the lesion core.

increased known axon-supportive substrate molecules¹¹, laminin, fibronectin and collagen in SCI lesions (Fig. 2a, Extended Data Fig. 5a), whereas potentially inhibitory chondroitin sulfate proteoglycans (CSPG)²⁸ were not significantly altered (Fig. 2g). FGF with EGF significantly increased astrocyte proliferation and density (Fig. 2b), yet despite this increase, stimulated and chemoattracted propriospinal axons regrow robustly through and beyond proximal (Fig. 2c) and distal (Fig. 3a) astrocyte scar borders, consistent with observations for sensory axons¹⁹. FGF with EGF also increased stromal cell density in the lesion core (Fig. 2a, Extended Data Fig. 6b). Axons transitioned readily from regrowing along astrocytes in proximal scar borders to regrowing along stromal cells in lesion core (Fig. 2c, Extended Data Fig. 6a), often orientated along stromal cells or blood vessels and circumventing

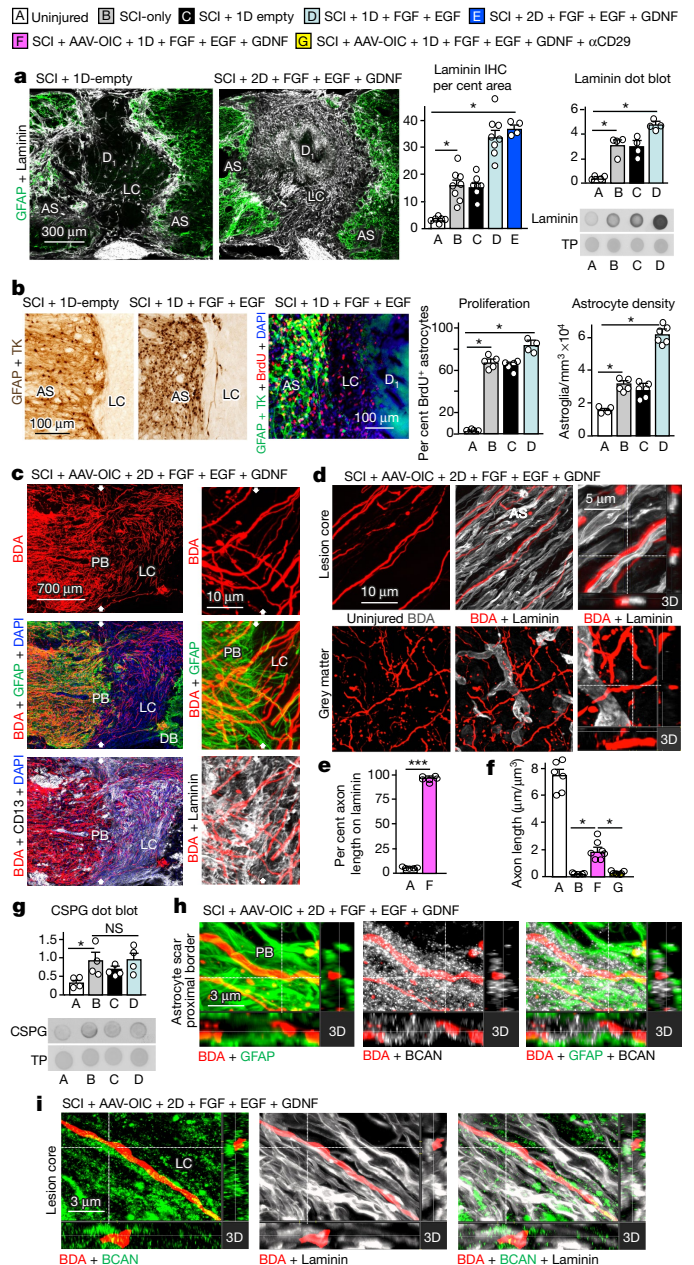


Fig. 2 | Stimulated, supported and chemoattracted mouse propriospinal axons regrow through the proximal borders of astrocyte scars and lesion core stromal cells along laminin that has been upregulated by delivered growth factors, in spite of CSPG presence. **a**, Laminin immunohistochemistry (IHC) images (left) plus quantification (middle, percentage stained area), and dot blot plus quantification of density (right). **P* < 0.01, one-way ANOVA with Bonferroni. TP, total protein. The colour key is used throughout all panels. **b**, IHC images and quantification (cell number) of astrocyte proliferation and density. **P* < 0.0005, one-way ANOVA with Bonferroni. **c**, BDA-labelled axon regrowth past proximal borders and in the lesion core among CD13⁺ stromal cells (left) and along laminin (right). White arrows denote proximal borders. **d–f**, IHC images (**d**), quantification of axon contact with laminin (**e**, ****P* < 0.0001; Student's two-tailed *t*-test, *t*(9) = 107.4) and axon length per tissue volume (**f**, **P* < 0.0005 one-way ANOVA with Bonferroni). **g**, CSPG dot blot and quantification of density. **P* < 0.05, one-way ANOVA with Bonferroni. For all graphs, data are mean ± s.e.m. and dots show *n* mice per group. **h**, **i**, BDA-labelled axon regrowth through astrocytes of proximal borders (**h**) and along laminin in the lesion centre (**i**) in spite of dense brevican (BCAN).

inflammatory cell clusters (Fig. 2c, Extended Data Fig. 6a–c). Some regrowing axons partially contacted cells expressing Schwann cell markers (Extended Data Fig. 6d). Thus, appropriately stimulated and

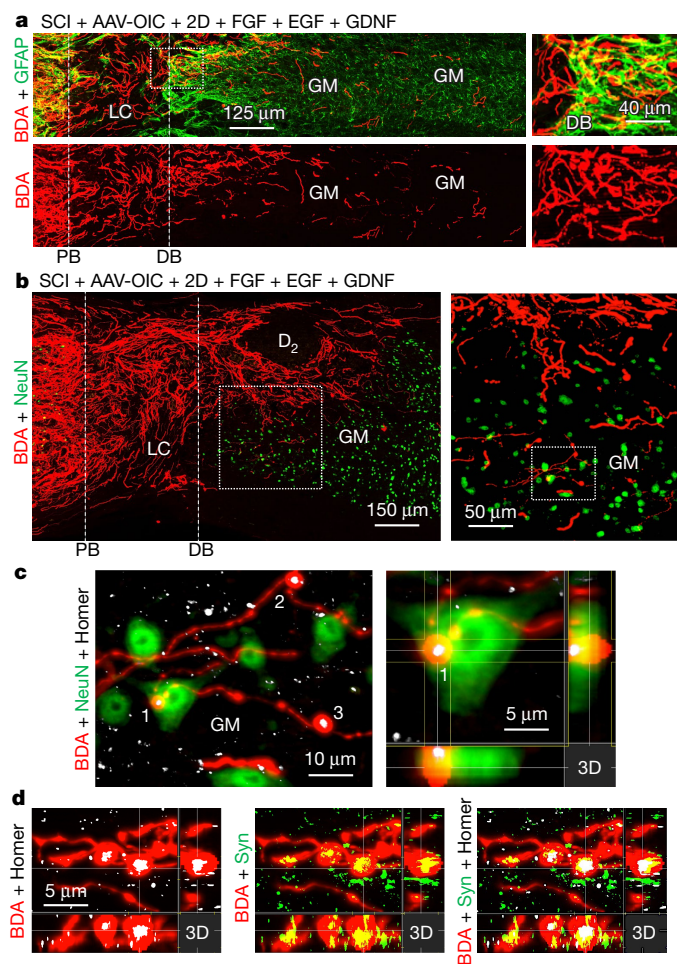


Fig. 3 | Stimulated and chemoattracted mouse propriospinal axons regrow past astrocyte scar distal borders into grey matter and form synapse-like contacts with neurons. **a**, **b**, Surveys and details (boxed areas) of BDA-labelled axon regrowth across distal borders and into grey matter. **c**, Detail of **b** (outlined region) and 3D view of synapse-like contact of BDA-labelled terminal with the post-synaptic marker, homer, on NeuN⁺ neuron. **d**, Synapse-like BDA-labelled terminals with overlapping pre- and post-synaptic markers, synaptophysin (Syn) and homer.

attracted axons regrew in contact with multiple cell types. Notably, over 98% of regrowing propriospinal axons in SCI lesions had at least one surface continually in contact with laminin, whereas axon surfaces in mature uninjured tissue rarely contacted laminin (Fig. 2d, e). Many regrowing axons also contacted fibronectin or collagen (Extended Data Fig. 5a). Simultaneous *in vivo* delivery of anti-CD29, an integrin-function-blocking antibody^{19,24}, significantly prevented most axon regrowth (Fig. 2f, Extended Data Figs. 3d, 4), demonstrating that regrowth required integrin-dependent axon–substrate interactions with laminin, fibronectin or collagen.

Nevertheless, upregulation of permissive substrate alone was not sufficient to attract activated axon regrowth. AAV-OIC plus depots of only FGF with EGF exhibited no significant regrowth, whereas AAV-OIC plus FGF, EGF and GDNF did (Fig. 1c, Extended Data Figs. 3d, 4), demonstrating that axon regrowth also required chemoattraction.

Notably, AAV-OIC plus FGF, EGF and GDNF stimulated axons regrew robustly through dense areas of CSPGs, including in direct contact with brevicin or CSPG4 (also known as NG2) in both astrocyte scars and non-neural lesion cores. This regrowth occurred along surfaces with high laminin expression (Fig. 2h, i, Extended Data Fig. 5b, c), consistent with *in vitro* observations that CSPG inhibition is relative rather than absolute, such that increasing laminin overrides CSPG presence²⁹.

To probe more broadly the effects of prolonged FGF and EGF treatment on astrocytes and other cells in SCI lesions, we conducted genome-wide sequencing of astrocyte-specific ribosome-associated RNA and RNA from non-astrocyte cells¹⁹ at two weeks after SCI. At this time point, FGF and EGF treatment continued to significantly regulate many genes. The most significantly regulated gene networks were associated with astrocyte proliferation and development, and with non-astrocyte inflammatory responses (Extended Data Fig. 7a–c).

We next identified mechanisms required to achieve propriospinal axon regrowth beyond lesion cores and distal borders into spared grey matter. In early experiments with one depot of AAV-OIC plus FGF, EGF and GDNF in lesion cores, we noted that propriospinal axons regrew robustly to surround and encircle depots, but did not pass beyond (Extended Data Fig. 4). We therefore injected a second depot of GDNF into spared grey matter caudal to injuries at one week after the first depot (Extended Data Fig. 1b). In mice with two such spatially and temporally separated depots, axons regrew robustly across lesion cores and distal astrocyte borders and routinely reached a full spinal segment past lesion centres (Fig. 1b–d, Extended Data Fig. 3a–c), demonstrating that chemoattraction is required to draw robust regrowth of mature endogenous axons into spared neural tissue beyond injuries. Notably, the second depot was placed at nine days after SCI, indicating that GDNF efficiently chemoattracted regrowing axons across already formed distal astrocyte scar borders without altering CSPG levels (Fig. 3a; Extended Data Fig. 3b). Starting just beyond lesion borders (Fig. 3b, c), propriospinal axons regrowing in spared grey matter intermingled with NeuN⁺ neurons, and some formed terminal-like swellings that contacted neurons, colocalized with the pre-synaptic marker synaptophysin and were in apposition with the post-synaptic marker homer (Fig. 3b–d). Such contacts were found wherever regrowing axons were present in grey matter, up to a full spinal segment (1,500 μ m) beyond lesion centres. As expected and discussed below, over-ground locomotion did not improve in these experiments focused on dissecting the mechanisms required to achieve axon regrowth across lesions (Extended Data Fig. 2f).

We next tested whether our findings could be extended to rats, in which lesion core pathophysiology has been proposed to be more similar to humans. As expected, rats exhibited little or no propriospinal axon regrowth after SCI with empty hydrogel. Axon regrowth was not increased by AAV-OIC alone, and only minimally by FGF, EGF and GDNF alone. By contrast, and consistent with our observations in mice, rats given combined AAV-OIC plus two depots of FGF, EGF and GDNF exhibited robust propriospinal axon regrowth that routinely reached a full spinal segment or more past lesion centres and penetrated well into spared grey matter around the second depot but not further (Fig. 4a, b, Extended Data Figs. 8, 9). Total axon regrowth past lesion centres was over 140-fold greater in rats with AAV-OIC plus FGF, EGF and GDNF versus SCI with empty hydrogel (Fig. 4b). AAV-derived tract-tracer, red fluorescent protein (RFP), was not expressed by axons of passage such as serotonin axons. Moreover, in contrast to the robust regrowth of RFP-labelled propriospinal axons, serotonin axons exhibited no regrowth (Extended Data Fig. 9c), indicating that our growth factor depots did not simply alter the lesion core environment to broadly enable regrowth of all axon types. Regrowing propriospinal axons that reached spared grey matter intermingled with NeuN⁺ neurons and some axons formed terminal-like swellings that contacted neurons, colocalized with synaptophysin and were in apposition with homer (Fig. 4c). As in mice, over-ground locomotion of rats did not improve in these experiments that probed the mechanisms required for axon regrowth but did not provide rehabilitation to elicit use-dependent plasticity (Extended Data Fig. 9d). Nevertheless, to look for potential basic functionality of regrown propriospinal axons, we measured electrophysiological signals across lesions. Rats with SCI only exhibited essentially no conduction above background levels across lesions, whereas rats with AAV-OIC plus FGF, EGF and GDNF exhibited conduction at about 25% of control levels at 2 mm past lesions, which disappeared by 5 mm past lesions (Fig. 4d), indicating that propriospinal

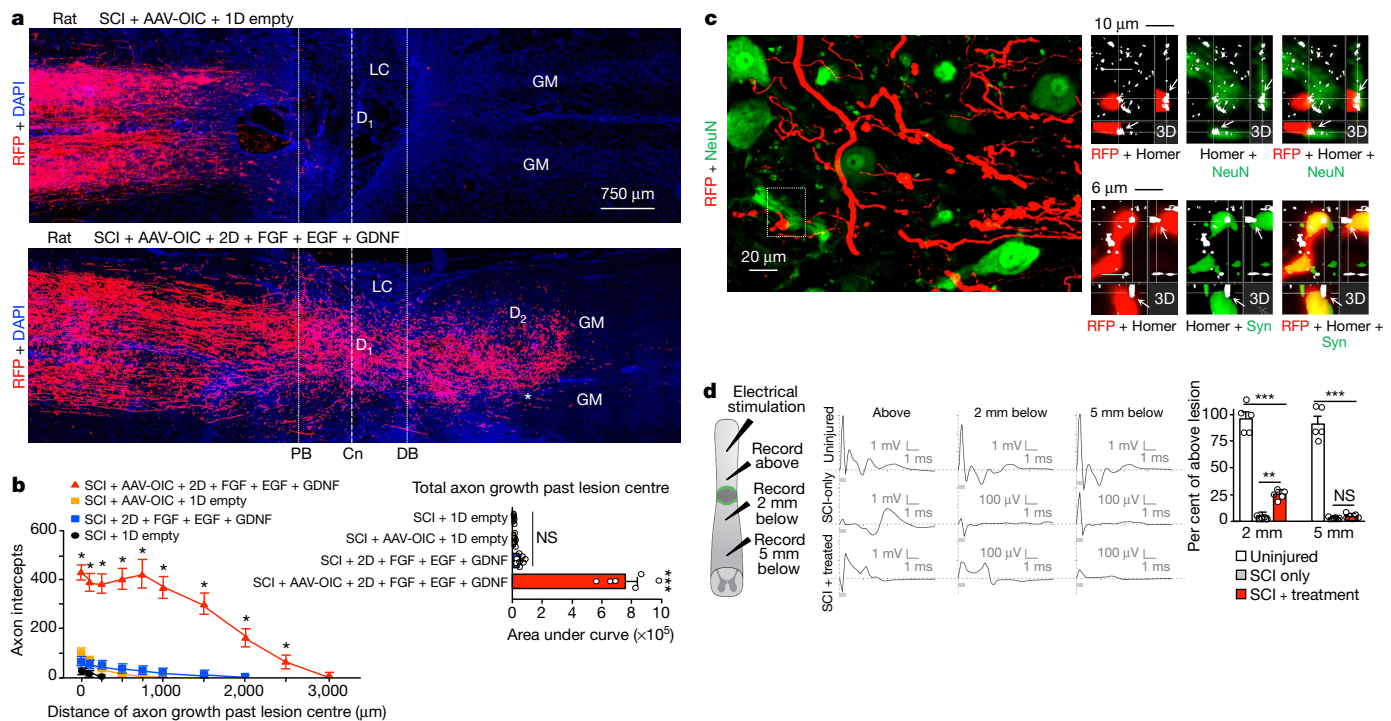


Fig. 4 | Stimulated and chemoattracted propriospinal axons regrow robustly and conduct electrophysiological signals across anatomically complete SCI lesions in rats after combined delivery of AAV-OIC plus FGF, EGF and GDNF in two sequentially placed hydrogel depots. **a**, RFP-labelled axons in composite tiled scans of horizontal sections. Dotted lines demarcate astrocyte proximal and distal borders around lesion core. Dashed line demarcates lesion centre (Cn). **b**, Left, axon intercepts at specific distances past lesion centres (colour coding and *n* as in bar graph). Right, areas under axon intercept curves. **P* < 0.01 versus all other groups, ****P* < 0.0001 versus all other groups, one-way ANOVA with Bonferroni. **c**, Detail images from the region indicated with an asterisk in **a**. Left, RFP-labelled axons among NeuN⁺ neurons in spared grey matter 2,000 μm past the lesion centre. Top right, 3D detail of the

axon regrowth was associated with a significant return of conduction capacity that correlated with the distance of regrowth past lesions.

Biological repair of anatomically complete SCI will require axon regrowth across lesions with non-neural tissue cores and astrocyte limitans borders to reach spared grey matter and form new circuits³. A mechanistic understanding of why spontaneous axon regrowth fails in adults is fundamental to creating beneficial interventions⁴. Our findings, in both mice and rats, strongly support the hypothesis that adult axon regrowth across such lesions fails primarily because of the simultaneous absence or inadequate presence of three types of mechanisms essential for facilitating developmental axon growth: (i) neuron intrinsic growth capacity^{2,5–9}, (ii) supportive substrate^{10,11} and (iii) chemoattraction^{12,13}. We show that each of these mechanisms is required, and that in combination—but not individually—they are sufficient to achieve robust axon regrowth across lesions in spite of the presence of putative growth inhibitors. We extend previous observations that providing individual growth-facilitators is not sufficient to achieve meaningful axon regrowth across complete lesions and combinations can improve regrowth^{3,4,19,23,30,31}. Importantly, our findings identify chemoattraction as critically required for robust axon regrowth, and point towards the need to identify chemoattractants effective for other axon populations desirable to target after SCI. The different response of propriospinal neurons to AAV-shPT and AAV-OIC reflects previous observations of neuron-specific activation requirements¹⁴ and underlines the need to identify growth activators for different neuronal populations and the means to achieve activation at subacute or chronic times after SCI. Although propriospinal axon regrowth was associated with a significant

return of electrophysiological conduction across lesions, there was, as expected⁴, no detectable improvement of locomotor function, consistent with accumulating evidence that new circuits formed after complete SCI cannot be expected to acquire function spontaneously, but will require rehabilitation that fosters their integration into functional networks through use-dependent plasticity^{4,20,21,32}. Our findings provide proof-of-concept evidence that robust and physiologically active descending propriospinal axon regrowth can be achieved across anatomically complete SCI lesions, and identify a mechanism-based biological repair strategy for such lesions that can be tested in conjunction with targeted rehabilitation paradigms³² designed to augment synapse remodelling and functional recovery of remodelling circuits.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0467-6>.

Received: 28 February 2018; Accepted: 13 July 2018;
Published online: 29 August 2018

1. Tessier-Lavigne, M. & Goodman, C. S. The molecular biology of axon guidance. *Science* **274**, 1123–1133 (1996).
2. He, Z. & Jin, Y. Intrinsic control of axon regeneration. *Neuron* **90**, 437–451 (2016).
3. O'Shea, T. M., Burda, J. E. & Sofroniew, M. V. Cell biology of spinal cord injury and repair. *J. Clin. Invest.* **127**, 3259–3270 (2017).
4. Sofroniew, M. V. Dissecting spinal cord regeneration. *Nature* **557**, 343–350 (2018).

5. Goldberg, J. L., Klassen, M. P., Hua, Y. & Barres, B. A. Amacrine-signaled loss of intrinsic axon growth ability by retinal ganglion cells. *Science* **296**, 1860–1864 (2002).
6. Bradke, F., Fawcett, J. W. & Spira, M. E. Assembly of a new growth cone after axotomy: the precursor to axon regeneration. *Nat. Rev. Neurosci.* **13**, 183–193 (2012).
7. Tedeschi, A. et al. The calcium channel subunit Alpha2delta2 suppresses axon regeneration in the adult CNS. *Neuron* **92**, 419–434 (2016).
8. Geoffroy, C. G., Hilton, B. J., Tetzlaff, W. & Zheng, B. Evidence for an age-dependent decline in axon regeneration in the adult mammalian central nervous system. *Cell Reports* **15**, 238–246 (2016).
9. Puttagunta, R. et al. PCAF-dependent epigenetic changes promote axonal regeneration in the central nervous system. *Nat. Commun.* **5**, 3527 (2014).
10. Letourneau, P. C. Cell-to-substratum adhesion and guidance of axonal elongation. *Dev. Biol.* **44**, 92–101 (1975).
11. Gundersen, R. W. Response of sensory neurites and growth cones to patterned substrata of laminin and fibronectin *in vitro*. *Dev. Biol.* **121**, 423–431 (1987).
12. Sperry, R. W. Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proc. Natl Acad. Sci. USA* **50**, 703–710 (1963).
13. Campenot, R. B. Local control of neurite development by nerve growth factor. *Proc. Natl Acad. Sci. USA* **74**, 4516–4519 (1977).
14. Duan, X. et al. Subtype-specific regeneration of retinal ganglion cells following axotomy: effects of osteopontin and mTOR signaling. *Neuron* **85**, 1244–1256 (2015).
15. Bei, F. et al. Restoration of visual function by enhancing conduction in regenerated axons. *Cell* **164**, 219–232 (2016).
16. Siebert, J. R., Middleton, F. A. & Stelzner, D. J. Intrinsic response of thoracic propriospinal neurons to axotomy. *BMC Neurosci.* **11**, 69 (2010).
17. Deng, L. X. et al. A novel growth-promoting pathway formed by GDNF-overexpressing Schwann cells promotes propriospinal axonal regeneration, synapse formation, and partial recovery of function after spinal cord injury. *J. Neurosci.* **33**, 5655–5667 (2013).
18. Nowak, A. P. et al. Rapidly recovering hydrogel scaffolds from self-assembling diblock copolypeptide amphiphiles. *Nature* **417**, 424–428 (2002).
19. Anderson, M. A. et al. Astrocyte scar formation aids central nervous system axon regeneration. *Nature* **532**, 195–200 (2016).
20. Courtine, G. et al. Recovery of supraspinal control of stepping via indirect propriospinal relay connections after spinal cord injury. *Nat. Med.* **14**, 69–74 (2008).
21. van den Brand, R. et al. Restoring voluntary control of locomotion after paralyzing spinal cord injury. *Science* **336**, 1182–1185 (2012).
22. Jacobi, A. et al. FGF22 signaling regulates synapse formation during post-injury remodeling of the spinal cord. *EMBO J.* **34**, 1231–1243 (2015).
23. Zukor, K. et al. Short hairpin RNA against PTEN enhances regenerative growth of corticospinal tract axons after spinal cord injury. *J. Neurosci.* **33**, 15350–15361 (2013).
24. Plantman, S. et al. Integrin-laminin interactions controlling neurite outgrowth from adult DRG neurons *in vitro*. *Mol. Cell. Neurosci.* **39**, 50–62 (2008).
25. Kashpur, O., LaPointe, D., Ambady, S., Ryder, E. F. & Dominko, T. FGF2-induced effects on transcriptome associated with regeneration competence in adult human fibroblasts. *BMC Genomics* **14**, 656 (2013).
26. White, R. E., Yin, F. Q. & Jakeman, L. B. TGF- α increases astrocyte invasion and promotes axonal growth into the lesion following spinal cord injury in mice. *Exp. Neurol.* **214**, 10–24 (2008).
27. Tuszynski, M. H. & Steward, O. Concepts and methods for the study of axonal regeneration in the CNS. *Neuron* **74**, 777–791 (2012).
28. Cregg, J. M. et al. Functional regeneration beyond the glial scar. *Exp. Neurol.* **253**, 197–207 (2014).
29. Tom, V. J., Steinmetz, M. P., Miller, J. H., Doller, C. M. & Silver, J. Studies on the development and behavior of the dystrophic growth cone, the hallmark of regeneration failure, in an *in vitro* model of the glial scar and after spinal cord injury. *J. Neurosci.* **24**, 6531–6539 (2004).
30. Richardson, P. M. & Issa, V. M. Peripheral injury enhances central regeneration of primary sensory neurones. *Nature* **309**, 791–793 (1984).
31. Alto, L. T. et al. Chemotropic guidance facilitates axonal regeneration and synapse formation after spinal cord injury. *Nat. Neurosci.* **12**, 1106–1113 (2009).
32. Asboth, L. et al. Cortico-reticulo-spinal circuit reorganization enables functional recovery after severe spinal cord contusion. *Nat. Neurosci.* **21**, 576–588 (2018).

Acknowledgements This work was supported by US National Institutes of Health (NS084030 to M.V.S., F32NS096858 to J.E.B., NS096294 to Z.H., and NS062691 to G.Cop.); Dr. Miriam and Sheldon G. Adelson Medical Foundation (M.V.S., Z.H., T.J.D. and G.Cop.); International Foundation for Research in Paraplegia (146 to M.A.A. and G.Cou.); ALARME Foundation (531066 to M.A.A. and G.Cou.); Association Song Taaba (M.A.A.); Craig H. Neilsen Foundation (381357 to T.M.O. and M.V.S.); Consolidator Grant from the European Research Council [ERC-2015-CoG HOW2WALKAGAIN 682999] (G.Cou.); Paralyzed Veterans Foundation of America (3080 to J.E.B. and M.V.S.); Swiss National Science Foundation (323530-164220 to S.L.B. and G.Cou.); Microscopy Core Resource of UCLA Broad Stem Cell Research Center; Microscopy Core Resource of the Wyss Center for Bio and Neuroengineering; and Wings for Life (M.V.S., J.E.B. and Z.H.).

Reviewer information *Nature* thanks J. Fawcett, P. Letourneau and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.A.A., T.M.O., J.E.B., T.J.D., Z.H., G.Cou. and M.V.S. designed experiments; M.A.A., T.M.O., J.E.B., Y.A., S.L.B., A.M.B., N.D.J., A.R., A.L.W. and C.W. conducted experiments; M.A.A., T.M.O., Y.A., J.E.B., N.D.J., J.H.K., B.K., R.K., G.Cop. and M.V.S. analysed data. M.A.A., T.M.O., J.E.B., T.J.D., G.Cou. and M.V.S. prepared the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0467-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0467-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to G.C. or M.V.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mice. All experiments using mice were conducted at UCLA using C57/BL6 female and male mice. RNA sequencing experiments used C57/BL6 mice expressing an mGFP-RiboTag transgene generated and characterized as described¹⁹. All mice used were young adults between ten weeks and four months old at the time of spinal cord injury. Mice receiving AAV injections were between six and nine weeks old at the time of AAV injection. All mice were housed in a 12-h light/dark cycle in a specific pathogen-free facility with controlled temperature and humidity and were allowed free access to food and water. Animal care, including manual bladder voiding, was performed at least twice daily or as needed after SCI for the duration of the experiment. All experiments were conducted according to protocols approved by the Animal Research Committee of the Office for Protection of Research Subjects at University of California Los Angeles.

Rats. All surgical procedures in rats were done at EPFL. Experiments were conducted on young adult female Lewis rats between two and four months of age (180–220 g body weight) housed three to a cage on a 12-h light/dark cycle with access to food and water *ad libitum*. Housing, surgery and euthanasia were performed in compliance with the Swiss Veterinary Law guidelines. Animal care, including manual bladder voiding, was performed twice daily after SCI for the duration of the experiment. All procedures and experiments were approved by the Veterinary Office of the canton of Vaud and the Veterinary Office of the canton of Geneva (Switzerland).

Surgical procedures for mice. All surgeries on mice were performed at UCLA under general anaesthesia with isoflurane in oxygen-enriched air using an operating microscope (Zeiss), and rodent stereotaxic apparatus (David Kopf). AAV injections were made two weeks before SCI to allow time for molecular expression and were targeted at propriospinal neurons between one and two segments rostral to the planned locations of SCI lesions after laminectomy of a single vertebra. AAV (see below) were injected into two sites (one on each side of the cord, 0.25 μ l (AAV2/9 OPN: 1×10^{13} , IGF: 5×10^{12} , CNTF: 5×10^{12} genome copies per ml in sterile saline)) 0.6 mm below the surface at 0.1 μ l per minute using glass micropipettes (ground to 50 to 100 μ m tips) connected via high-pressure tubing (Kopf) to 10- μ l syringes under the control of microinfusion pumps. Severe crush SCIs were made at the level of T10 after laminectomy of a single vertebra by using No. 5 Dumont forceps (Fine Science Tools) without spacers and with a tip width of 0.5 mm to completely compress the entire spinal cord laterally from both sides for 5 s^{19,33–35} (Extended Data Fig. 1). Hydrogel depots were injected stereotactically into the centre of SCI lesions 0.6 mm below surface at 0.15 μ l per minute using glass micropipettes (ground to 50- to 100- μ m tips) connected via high-pressure tubing (Kopf) to 10- μ l syringes under the control of microinfusion pumps, 2 days after SCI³⁶. Tract-tracing of propriospinal neurons was performed by injection of biotinylated dextran amine 10000 (BDA, Sigma) 10% w/v in sterile saline injected $2 \times 0.4 \mu$ l into the same rostral segments targeted with AAV injections as described above. In animals receiving two hydrogel depots, the second depot was placed 1 mm caudal to the SCI 9 days after SCI. Timelines of all injections are provided in Extended Data Fig. 1. All mice received analgesic before wound closure and every 12 h for at least 48 h after injury. Animals were randomly assigned numbers and thereafter were evaluated blind to experimental condition.

Surgical procedures for rats. All surgeries on rats were performed at EPFL under general anaesthesia with isoflurane in oxygen-enriched air using an operating microscope (Zeiss), and rodent stereotaxic apparatus (David Kopf). AAV injections were made two weeks before SCI to allow time for molecular expression and were targeted at propriospinal neurons one and two segments rostral to planned locations of SCI lesions after laminectomy of a single vertebra. AAV were injected into four sites (two on each side of the cord, 0.25 μ l (AAV2/9 OPN: 1×10^{13} , IGF: 5×10^{12} , CNTF: 5×10^{12} genome copies per ml in sterile saline)) 1.1 mm below the surface at 0.2 μ l per minute using glass micropipettes connected via high pressure tubing (Kopf) to 10 μ l syringes under the control of microinfusion pumps. Severe crush SCI was made at the level of T10 laminectomy of a single vertebra by using No. 2 Dumont Forceps (Fine Science Tools) without spacers and with a tip of 0.5 mm to completely compress the entire spinal cord laterally from both sides for 5 s (Extended Data Fig. 1). Hydrogel depots were injected stereotactically into the centre of SCI lesions 1.1 mm below the surface at 0.2 μ l per minute using glass micropipettes connected via high pressure tubing (Kopf) to 10 μ l syringes under the control of microinfusion pumps, 2 days after SCI. One week later, a hydrogel depot was placed 2 mm caudal to the SCI. During the same surgery, tract-tracing of propriospinal neurons was performed by injection of AAV2/5 red fluorescent protein (RFP, University of Pennsylvania Vector Core, 2.612×10^{13} genome copies per ml) injected $4 \times 0.25 \mu$ l into the same rostral segments targeted with AAV injections as described above. Timelines of all injections are provided in Extended Data Fig. 1. All rats received analgesia (buprenorphine Temgesic, ESSEX Chemie AG, Switzerland, 0.01–0.05 mg per kg, subcutaneously) and antibiotics (Baytril 2.5%, Bayer Health Care AG, 5–10 mg per kg, subcutaneously) were provided for

3 and 5 days after surgery, respectively. Animals were randomly assigned numbers and thereafter were evaluated blind to experimental condition.

AAVs. Various AAVs were used to deliver either control AAV of a nonsense scrambled sequence for green fluorescent protein (AAV2/1-scrambled: 5×10^{12} genome copies per ml) or PTEN knockdown (AAV2/1-shPTEN: 5×10^{12} genome copies per ml)²³; or to express the growth factors osteopontin (OPN), IGF1 and CNTF (AAV2/9 OPN: 1×10^{13} genome copies per ml; AAV2/9 IGF-1: 5×10^{12} genome copies per ml; AAV2/9 CNTF: 5×10^{12} genome copies per ml)^{14,15}; or to express green fluorescent protein (GFP) as a reporter protein (AAV2/9 GFP: 2×10^{13} genome copies per ml) or red fluorescent protein (RFP) as an axonal tract-tracer (AAV2/5-RFP: 2.612×10^{13} genome copies per ml) (University of Pennsylvania Vector Core).

Hydrogel depots with growth factors and function-blocking antibodies. Biomaterial depots were prepared using well-characterized diblock copolypeptide hydrogels that are CNS biocompatible, biodegrade over several weeks in vivo and provide prolonged delivery of bioactive growth factors in CNS tissue for two or more weeks after injection^{18,19,37,38}. Diblock copolypeptide hydrogel K₁₈₀L₂₀ was fabricated, conjugated with blue fluorescent dye (AMCA-X) and loaded with growth factor and antibody cargoes as described^{36–38}. Cargo molecules were as follows. Human recombinant FGF2, EGF and GDNF were purchased from PeproTech: (i) human FGF2 (FGF-basic) (154 amino acids) Cat#100-18B-100UG, Lot#091608 C0617; (ii) human EGF Cat#AF-100-15-100UG, Lot#0816AFC05 B2317; (iii) human GDNF Cat#405-10-100UG, Lot#0606B64 A2517. Integrin-function-blocking hamster anti-rat CD29 monoclonal antibody was purchased from BD Bioscience as a custom order at 10.3 mg/ml (product #624084; lot#7165896). Freeze-dried K₁₈₀L₂₀ powder was reconstituted to 3.0% or 3.5% w/v in sterile PBS without cargo or with combinations of FGF2 (1.0 μ g/ μ l), EGF (1.0 μ g/ μ l), GDNF (1.0 μ g/ μ l) and anti-CD29 (5 μ g/ μ l). Diblock copolypeptide hydrogel formulations were prepared to have G' (storage modulus measured at 10 rad/s and 10% strain) between 75 and 100 Pa, somewhat below that of mouse brain at 200 Pa^{37,38}.

Hindlimb locomotor evaluation. At 2, 7, 14 and 28 days after SCI, hindlimb movements were scored using a simple six-point scale in which 0 is no movement and 5 is normal walking³⁴.

Animal inclusion and exclusion criteria. Two days after SCI, all mice or rats were evaluated in open field and all animals exhibiting any hindlimb movements were not studied further. Rodents that passed this inclusion criterion were randomized into experimental groups for further treatments and were thereafter evaluated blind to their experimental condition.

Histology and immunohistochemistry. After terminal anaesthesia by barbiturate overdose mice or rats were perfused transcardially with 4% paraformaldehyde and spinal cords processed for immunofluorescence as described^{19,33–35}. Primary antibodies were: rabbit anti-GFAP (1:2,000; Dako); rat anti-GFAP (1:1,000, Thermofisher); chicken anti-GFAP (1:1,000, Novus Biologicals); rabbit anti-NeuN (1:1,000, Abcam); rabbit anti-GDNFR- α (GDNF-receptor α) (1:1,000, Abcam); sheep anti-BrdU (1:300, Maine Biotechnology Services); rabbit anti-HSV-TK (1:1,000^{35,39}); goat anti-CD13 (1:1,000, R&D systems); rabbit anti-laminin 1 (1:100, Sigma); rabbit anti-fibronectin (1:500, Millipore); rabbit anti-collagen 1a1 (1:300, Novus Biologicals); mouse anti-NeuN (1:2,000, Millipore); mouse anti-CSPG⁴⁰ (1:100, Sigma); rabbit anti-brevican (BCAN) (1:300, Novus Biologicals); guinea pig anti-NG2 (CSPG4) (E. G. Hughes and D. W. Bergles⁴¹); rat anti-PECAM-1 (1:200, BD Biosciences); guinea pig anti-homer1 (1:600, Synaptic Systems GmbH); rabbit anti-synaptophysin (1:600, Dako); rabbit anti-RFP (1:1,000, Rockland); chicken anti-RFP (1:500, Novus Biologicals); goat anti-GFP (1:1,000, Novus Biologicals). Fluorescence secondary antibodies were conjugated to: Alexa 488 (green) or Alexa 405 (blue) (Molecular Probes), or to Cy3 (550, red) or Cy5 (649, far red) all from (Jackson ImmunoResearch Laboratories). BDA tract-tracing was visualized with streptavidin–horse radish peroxidase (HRP) plus TSB Fluorescein green or Tyr-Cy3 (Perkin Elmer). Nuclear stain: 4',6'-diamidino-2-phenylindole dihydrochloride (DAPI; 2 ng/ml; Molecular Probes). Sections were coverslipped using ProLong Gold anti-fade reagent (Invitrogen). Sections were examined and photographed using deconvolution fluorescence microscopy and scanning confocal laser microscopy (Zeiss). Tiled scans of individual whole sections were prepared using a 20 \times objective and the scanning function of a Leica Aperio Versa 200 Microscope (Leica) available in the UCLA Translational Pathology Core Laboratory. Composite survey images were prepared from tiled scans of multiple sections from the same animals oriented and overlaid using Imaris software (9.1.2.64 Bit, Bitplane, Oxford Instruments).

Axon quantification. Axons labelled by tract tracing using BDA or RFP were quantified using image analysis software (NeuroLucida, 9.14.5.32 Bit, MicroBrightField) operating a computer-driven microscope regulated in the x, y and z axes (Zeiss) by observers blind to experimental conditions. Using NeuroLucida, lines were drawn across horizontal spinal cord sections at SCI lesion centres and at regular distances beyond (Fig. 1c, Extended Data Fig. 3a, b) and the number of axons intercepting lines was counted by observers blind to experimental conditions. Multiple sections

through the middle of the cord, in which propriospinal axons were densest, were counted per mouse or rat and expressed as total intercepts per location per animal. To determine the efficacy of axon transection after SCI, we examined labelling 3 mm distal to SCI lesion centres in mice and 5 mm distal to lesion centres in rats, with the intention of eliminating animals that had labelled axons at this location on the grounds that these mice may have had incomplete lesions. However, essentially all mice or rats that had met the strict behavioural inclusion criterion of no hindlimb movements 2 days after severe crush SCI exhibited no detectable axons 3 mm or 5 mm, respectively, distal to SCI lesions, regardless of treatment group.

Quantification of immunohistochemically stained areas. Sections stained for laminin 1, fibronectin, collagen 1a1 or CSPG were scanned using constant exposure settings. Single channel immunofluorescence images were converted to black and white and thresholded (Fig. 2a) and the amount of stained area was measured in different tissue compartments using NIH ImageJ (v.1.51) software¹⁹.

Quantification of astrocyte proliferation and density. To quantify astrocyte proliferation and the number of astrocytes in the immediate scar border previously defined as zone 1³⁵, we used a well-characterized transgenic mouse line that expresses thymidine kinase in astrocyte cell bodies, thereby facilitating quantification of cell number and co-localization with other markers³⁵. To quantify the proportion of newly proliferated astroglia in the scar border, we injected daily single doses of the cell division marker, bromodeoxyuridine (BrdU, Sigma), 100 mg/kg/day dissolved in saline plus 0.007 N NaOH on days 2–7 after SCI. Newly proliferated astrocytes were quantified by determining the percentage of astrocytes stained for both GFAP, thymidine kinase and BrdU in zone 1³⁵. Total astrocyte numbers in the immediate scar border (zone 1)³⁵ were determined by counting the number of cells per defined tissue volume. Cell counts were performed using stereological image analysis software (StereoInvestigator, 9.14.5 32 Bit, and NeuroLucida, 9.14.5 32 Bit, MicroBrightField) operating a computer-driven microscope regulated in the *x*, *y* and *z* axes (Zeiss).

Dot blot. For the dot blot immunoassay of laminin 1, fibronectin, collagen 1 or CSPG, spinal cord tissue blocks were lysed and homogenized in standard RIPA (radio-immuno-precipitation-assay) buffer. LDS (lithium dodecyl sulfate) buffer (Life Technologies) was added to the post-mitochondrial supernatant and 2 µl containing 2 µg/µl protein was spotted onto a nitrocellulose membrane (Life Technologies), set to dry and incubated overnight with primary antibodies: rabbit anti-laminin 1 (1:4,000, Sigma); rabbit anti-fibronectin (1:7,000, Millipore); rabbit anti-collagen 1a1 (1:7,000, Novus Biologicals); mouse anti-chondroitin sulfate antibody (CS-56, 1:3,000, Sigma Aldrich), an IgM-monoclonal antibody that detects glyco-moiety of all CSPGs⁴⁰. Immunoreactivity was detected on X-ray film with HRP-conjugated secondary antibody (1:5,000) and chemiluminescent substrate (Thermo Fisher). Densitometry measurements of immunoreactivity were obtained using ImageJ software (NIH) and normalized to total protein (Ponceau S) density⁴². Raw images of dot blots are provided as Supplementary Fig. 1.

Isolation and sequencing of RNA from astrocytes and non-astrocyte cells. Using mice expressing an mGFAP-RiboTag transgene, RNA was evaluated as previously described¹⁹ from uninjured mice, and mice two weeks after SCI after treatment with hydrogel depots that either contained no cargo (empty depots) or delivered FGF + EGF (Extended Data Fig. 1a). In brief, spinal cords were rapidly dissected out of the spinal canal and the central 3 mm of the lower thoracic lesion including the lesion core and 1 mm rostral and caudal were rapidly removed and snap-frozen in liquid nitrogen. Haemagglutinin immunoprecipitation of astrocyte ribosomes and ribosome-associate mRNA was carried out as described⁴³. The non-precipitated flow through from each sample was collected for analysis of non-astrocyte total RNA. Haemagglutinin and flow-through samples underwent on-column DNA digestion using the RNase-Free DNase Set (Qiagen) and RNA purified with the RNeasy Plus Micro kit (Qiagen). Integrity of the eluted RNA was analysed using a 2100 Bioanalyzer (Agilent) with the RNA Pico chip, average RNA integrity number (RIN) = 7.9 ± 1.4 . RNA concentration determined using the RiboGreen RNA Assay kit (Life Technologies). cDNA was generated from 5 ng of immunoprecipitate or flow-through RNA using the Nugen Ovation 2 RNA-Seq System V2 kit (Nugen). One microgram of cDNA was fragmented using the Covaris M220. Paired-end libraries for multiplex sequencing were generated from 300 ng of fragmented cDNA using the Apollo 324 automated library preparation system (Wafergen Biosystems), enriched over ten cycles of PCR and purified with Agencourt AMPure XP beads (Beckman Coulter). All samples were analysed by an Illumina NextSeq 500 Sequencer (Illumina) using 75-bp pair-end sequencing. The number of reads obtained are between 38.0 to 71.0 million (average 52.1 million). Sequences were aligned to mouse mm10 genome using STAR aligner (v.2.4.0j). Uniquely aligned reads were between 57.1 to 87.0% (average 73.9%). Read counts were determined using HT-seq (v.0.6.0). Differential expression analysis was conducted using Bioconductor EdgeR package (v.3.20.1) after removal of low count genes (five counts for at least two samples). Three samples from astrocyte samples were excluded owing to low RIN values (<5.7), no samples were excluded for non-astrocyte samples.

Rat electrophysiology. Terminal electrophysiological assessments were carried out as previously described⁴⁴ four weeks after SCI. In brief, animals were anaesthetized with urethane (1.5 g/kg; intraperitoneally) and core body temperature was maintained at 37 °C using a self-regulating heated pad connected to a rectal probe. Depth of anaesthesia was continually monitored by assessing withdrawal reflexes and respiratory rate. After laminectomy to expose the injury site and 7 mm rostral and caudal to the lesion site by removal of one vertebra on either side of the injury, the dura was removed, and the exposed spinal cord was covered with warm mineral oil to prevent drying. For stimulation, a tungsten bipolar concentric electrode was positioned intraspinally at the rostral-most point of the laminectomy. A silver ball electrode was used to record any evoked activity from the surface of the exposed spinal cord at various locations (above, 2 mm below, and 5 mm below the lesion). Stimulation was delivered in 200 µs square wave pulses at the maximum amplitude possible before large motor responses were evoked (typically between 600 µA and 800 µA) and at a frequency of 0.75 Hz using a STG 4004 stimulus generator (Multi Channel Systems). Evoked activity was amplified and recorded using an A-M systems differential amplifier, PowerLab and LabChart Pro acquisition and analysis system (AD Instruments). For analysis, 30 traces from each recording site were averaged and the peak to peak amplitude of the evoked potential was quantified.

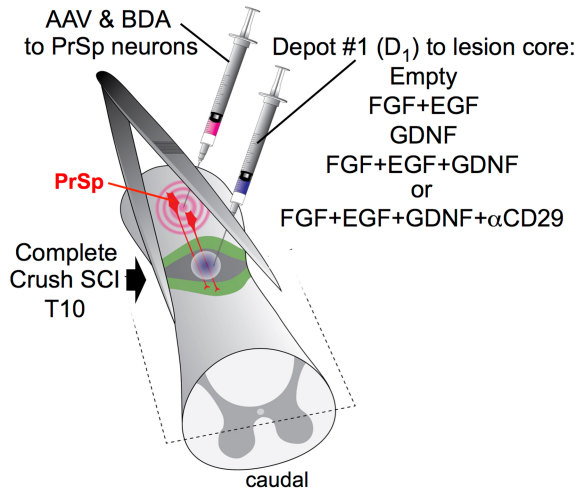
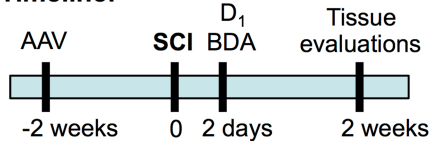
Statistics, power calculations, group sizes and reproducibility. Statistical evaluations of repeated measures were conducted by one-way ANOVA with post hoc independent pair-wise analysis as per Bonferroni, or by Student's *t*-test (Prism, 7.0c, GraphPad). For one-way ANOVA statistical evaluations, *F* values are also reported in the online Source Data files in the format *F*(degree of freedom 1, degree of freedom 2) = *X*. The degrees of freedom are computed as degree of freedom 1 = *k* − 1, in which *k* is the number of compared treatments, and degree of freedom 2 = *n* − *k* in which *n* is the total number of samples across the treatment groups. For Student's two-tailed *t*-test (Prism, 7.0c, GraphPad), *t* value and degrees of freedom are reported in the format *t*(degree of freedom) = *X*. Power calculations were performed using G*Power Software v.3.1.9.2⁴⁵. For quantification of histologically derived neuroanatomical outcomes such as numbers of axons or percentage of area stained, group sizes were used that were calculated to provide at least 80% power when using the following parameters: probability of type I error (α) = 0.05, a conservative effect size of 0.25, 3–10 treatment groups with multiple measurements obtained per replicate. All graphs show mean \pm s.e.m. as well as individual values as dot plots. All bar graphs are overlaid with dot plots in which each dot represents the value for one animal to show the distribution of data and the number (*n*) of animals per group. Files of all individual values are provided as Source Data. The main experiments testing propriospinal axon regrowth across SCI lesions in animals treated with SCI + AAV-OIC + 2D + FGF + EGF + GDNF and the main control groups (SCI + 1D empty, SCI + AAV-OIC + 1D empty, SCI + 2D + FGF + EGF + GDNF) were repeated independently three times in different groups of mice and three times in different groups of rats with similar results. Other experiments testing propriospinal axon regrowth across SCI lesions in animals in all other groups were repeated independently at least twice in different groups of mice with similar results. For all photomicrographs of histological tissue, staining experiments were repeated independently with tissue from at least four, and in most cases six, different animals with similar results.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

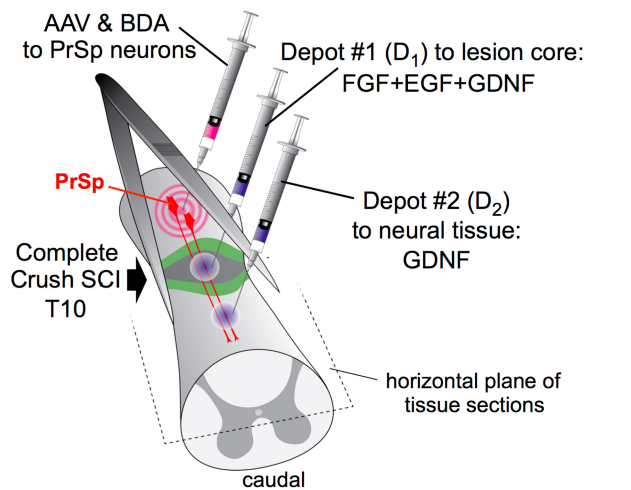
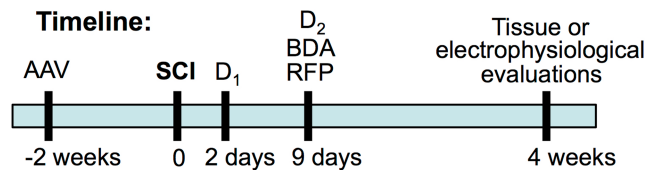
Data availability. Files of Source Data of individual values for all quantitative figures are provided with the paper. Raw images of dot blots are provided as Supplementary Fig. 1. RNA-seq data are available at the NCBI Gene Expression Omnibus under accession number GSE111529. Other data that support the findings of this study are available from the corresponding authors upon reasonable request.

33. Faulkner, J. R. et al. Reactive astrocytes protect tissue and preserve function after spinal cord injury. *J. Neurosci.* **24**, 2143–2155 (2004).
34. Herrmann, J. E. et al. STAT3 is a critical regulator of astrogliosis and scar formation after spinal cord injury. *J. Neurosci.* **28**, 7231–7243 (2008).
35. Wanner, I. B. et al. Glial scar borders are formed by newly proliferated, elongated astrocytes that interact to corral inflammatory and fibrotic cells via STAT3-dependent mechanisms after spinal cord injury. *J. Neurosci.* **33**, 12870–12886 (2013).
36. Zhang, S. et al. Tunable diblock copolypeptide hydrogel depots for local delivery of hydrophobic molecules in healthy and injured central nervous system. *Biomaterials* **35**, 1989–2000 (2014).
37. Yang, C. Y. et al. Biocompatibility of amphiphilic diblock copolypeptide hydrogels in the central nervous system. *Biomaterials* **30**, 2881–2898 (2009).
38. Song, B. et al. Sustained local delivery of bioactive nerve growth factor in the central nervous system via tunable diblock copolypeptide hydrogel depots. *Biomaterials* **33**, 9105–9116 (2012).
39. Bush, T. G. et al. Leukocyte infiltration, neuronal degeneration, and neurite outgrowth after ablation of scar-forming, reactive astrocytes in adult transgenic mice. *Neuron* **23**, 297–308 (1999).

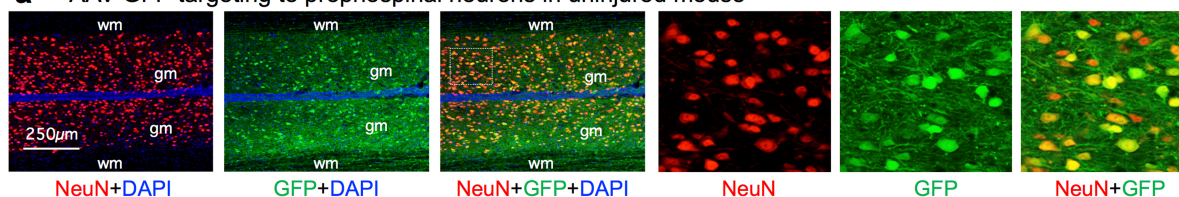
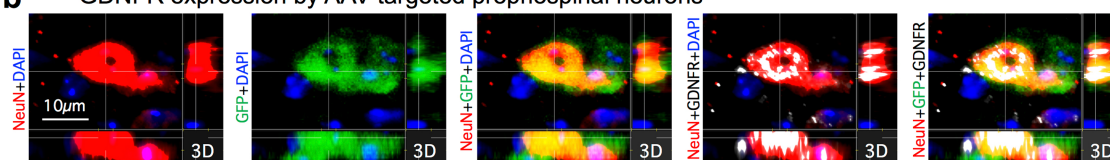
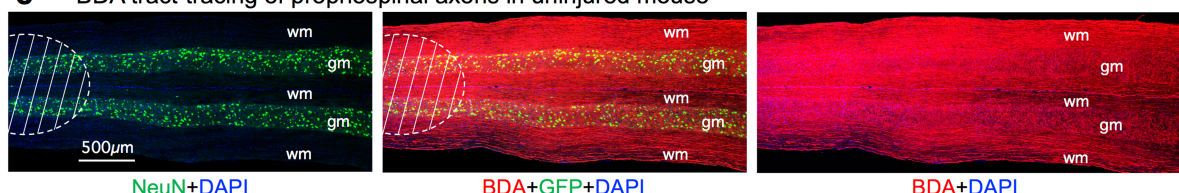
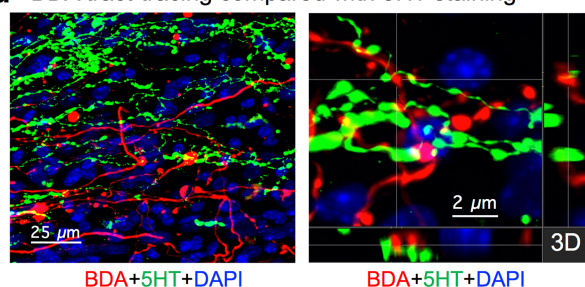
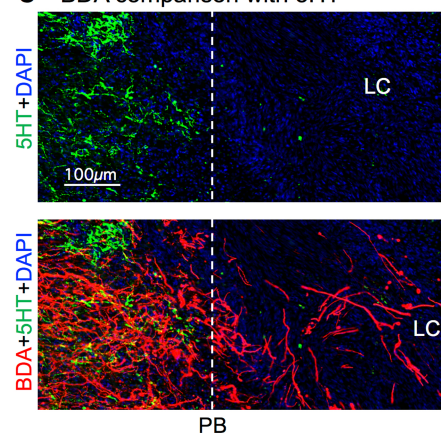
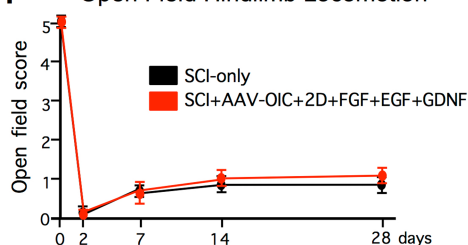
40. Avnur, Z. & Geiger, B. Immunocytochemical localization of native chondroitin-sulfate in tissues and cultured cells using specific monoclonal antibody. *Cell* **38**, 811–822 (1984).
41. Hughes, E. G., Kang, S. H., Fukaya, M. & Bergles, D. E. Oligodendrocyte progenitors balance growth with self-repulsion to achieve homeostasis in the adult brain. *Nat. Neurosci.* **16**, 668–676 (2013).
42. Romero-Calvo, I. et al. Reversible Ponceau staining as a loading control alternative to actin in western blots. *Anal. Biochem.* **401**, 318–320 (2010).
43. Sanz, E. et al. Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. *Proc. Natl Acad. Sci. USA* **106**, 13939–13944 (2009).
44. James, N. D. et al. Conduction failure following spinal cord injury: functional and anatomical changes from acute to chronic stages. *J. Neurosci.* **31**, 18543–18555 (2011).
45. Faul, F., Erdfelder, E., Lang, A. G. & Buchner, A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191 (2007).
46. Harrison, M. et al. Vertebral landmarks for the identification of spinal cord segments in the mouse. *Neuroimage* **68**, 22–29 (2013).
47. Watson, C., Paxinos, G. & Kayalioglu, G. *The Spinal Cord*. (Elsevier, London, 2008).

a Experimental model – 1 Depot (1D)**Timeline:**

Extended Data Fig. 1 | Experimental models and timelines. Mice or rats received different combinations of procedures including adeno-associated virus (AAV) injections, complete crush SCI, injections of one or two depots of hydrogel containing different molecular cargoes and injections of biotinylated dextran amine (BDA) for axonal tract-tracing. AAV injections were made two weeks before SCI to allow time for molecular expression and were targeted at propriospinal neurons (PrSp) between one and two segments rostral to planned locations of SCI lesions. AAVs were used to deliver either potential axon-growth reactivating molecules, GFP to identify targeted neurons or RFP as an axonal tract-tracer. Complete crush SCI lesions were placed at the level of spinal segment T10. Two days after SCI, all animals were behaviourally evaluated for completeness of SCI and only animals with functionally complete SCI were included in subsequent experimental steps. Additional animals with complete SCI were evaluated without hydrogel injections (SCI only). **a**, Schematic and timeline of one-depot experiments. Two days after complete crush

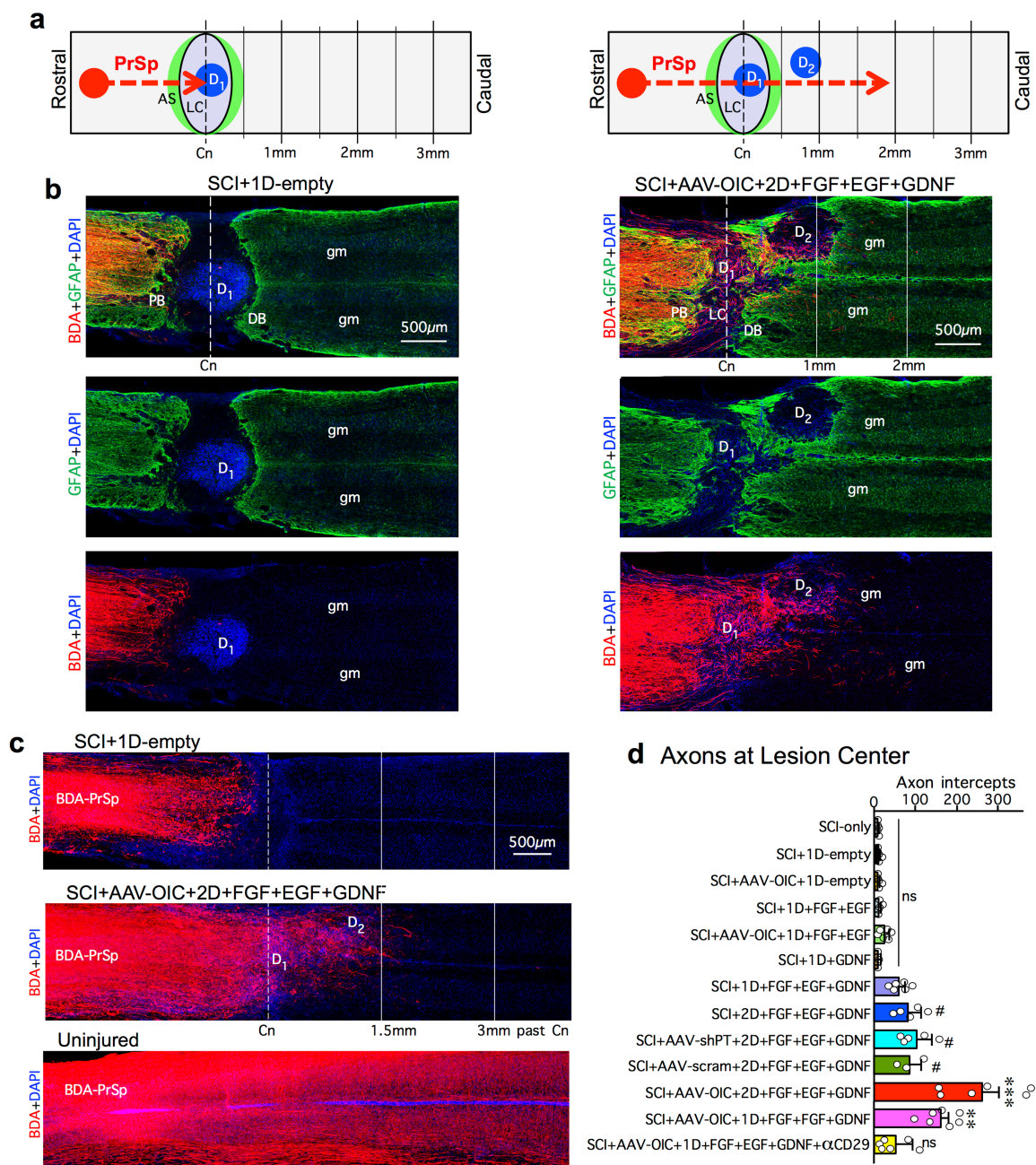
b Experimental model – 2 Depots (2D)**Timeline:**

SCI, animals received hydrogel injections targeted to the centre of the non-neural lesion core. These depots (D_1) contained different molecular cargoes as listed in the schematic. Depots without cargo were referred to as 'empty'. **b**, Schematic and timeline of two-depot experiments. Two days after complete crush SCI, animals received a D_1 hydrogel injection into the centre of the non-neural lesion core to deliver the growth factors FGF + EGF + GDNF. Nine days after SCI, the animals received a second hydrogel injection (D_2) targeted to spared neural tissue 1 to 2 mm caudal to the lesion centre to deliver GDNF to sequentially chemoattract propriospinal axons that had regrown into the lesion core. BDA injections for axonal tract-tracing were targeted at propriospinal neurons between one and two segments rostral to SCI lesions and were placed at the time of injecting either D_1 (**a**) or D_2 (**b**). Tissue was collected for evaluation at either two or four weeks after SCI. Electrophysiological evaluations were conducted at four weeks after SCI. For abbreviations, see Extended Data Table 1.

a AAV-GFP targeting to propriospinal neurons in uninjured mouse**b** GDNFR expression by AAV-targeted propriospinal neurons**c** BDA tract-tracing of propriospinal axons in uninjured mouse**d** BDA tract-tracing compared with 5HT staining**e** BDA comparison with 5HT**f** Open Field Hindlimb Locomotion

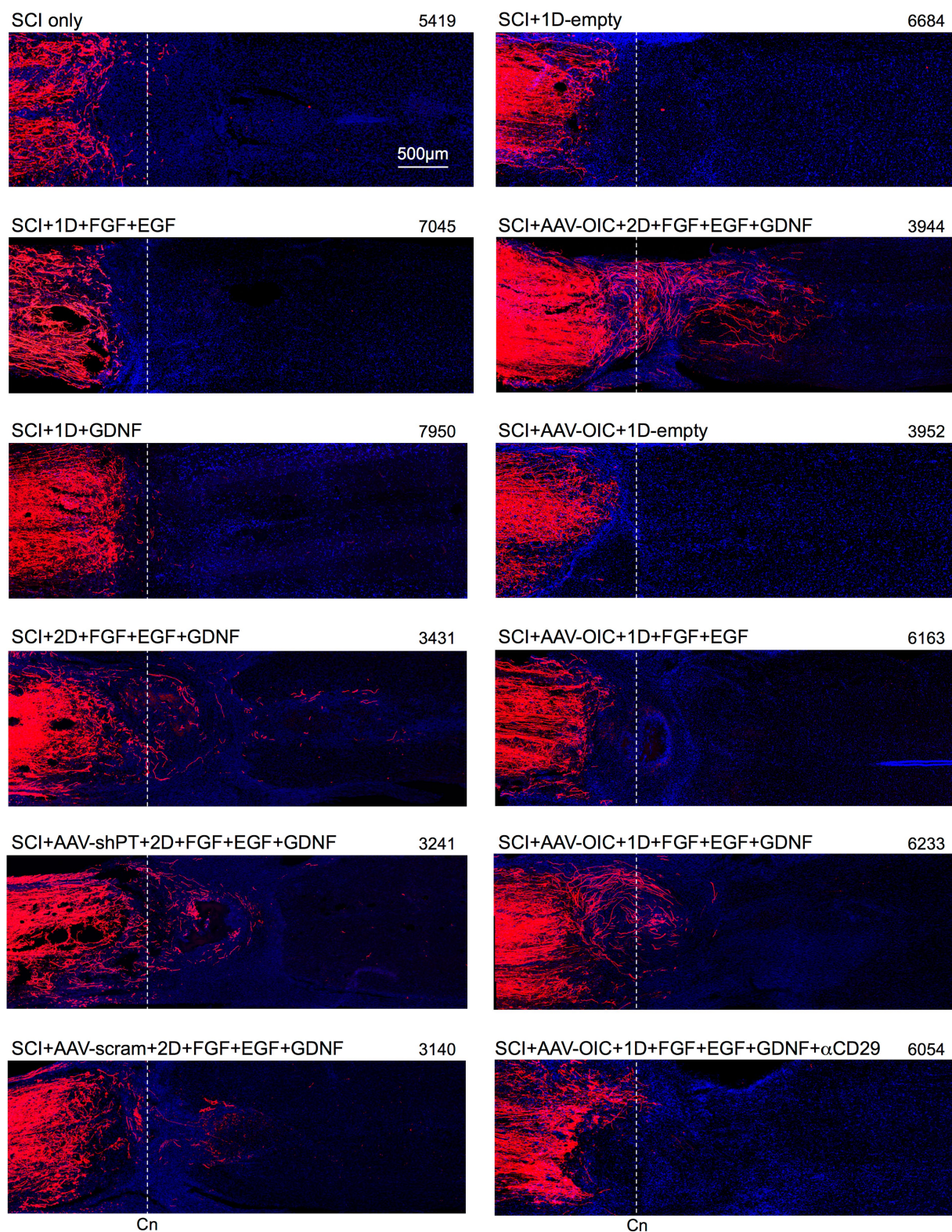
Extended Data Fig. 2 | AAV targeting, axon tracing and axon quantification. **a**, AAV targeting of green fluorescent protein (GFP) to propriospinal neurons. Multi-fluorescent, survey (left) and detail (right, boxed area) confocal images of horizontal section through mouse grey (gm) and white (wm) matter. Essentially all NeuN⁺ propriospinal neurons targeted with AAV express GFP. **b**, Multi-fluorescent, orthogonal 3D confocal images show that AAV-targeted propriospinal neurons express GDNFR. **c**, Multi-fluorescent, survey images show tract-tracing of propriospinal axons using biotinylated dextran amine (BDA) in tiled confocal scans of horizontal section from uninjured mouse. Hatched area indicates densely labelled location of BDA injections. **d**, **e**, Multiple channel fluorescent images compare BDA-labelled propriospinal axons and immunohistochemically stained serotonin (5HT) axons in

mice after SCI + AAV-OIC + 1D + FGF + EGF + GDNF. **d**, Survey and orthogonal 3D confocal detail from an area proximal to the SCI lesion shows a complete lack of overlap of BDA-labelling and 5HT immunohistochemistry, indicating that BDA-tracing did not label 5HT axons of passage. **e**, Survey images of the same field examined with different filters show BDA-labelled propriospinal axons (bottom image) regrowing robustly past the astrocyte scar proximal border and through the non-neural lesion core; by contrast, 5HT axons (top and bottom image) did not regrow into or through the lesion core. **f**, Open field hindlimb locomotor score at various times after SCI assessed using a 6-point scale in which 5 is normal walking and 0 is no movement of any kind. Data are mean \pm s.e.m., $n = 6$ mice per group.



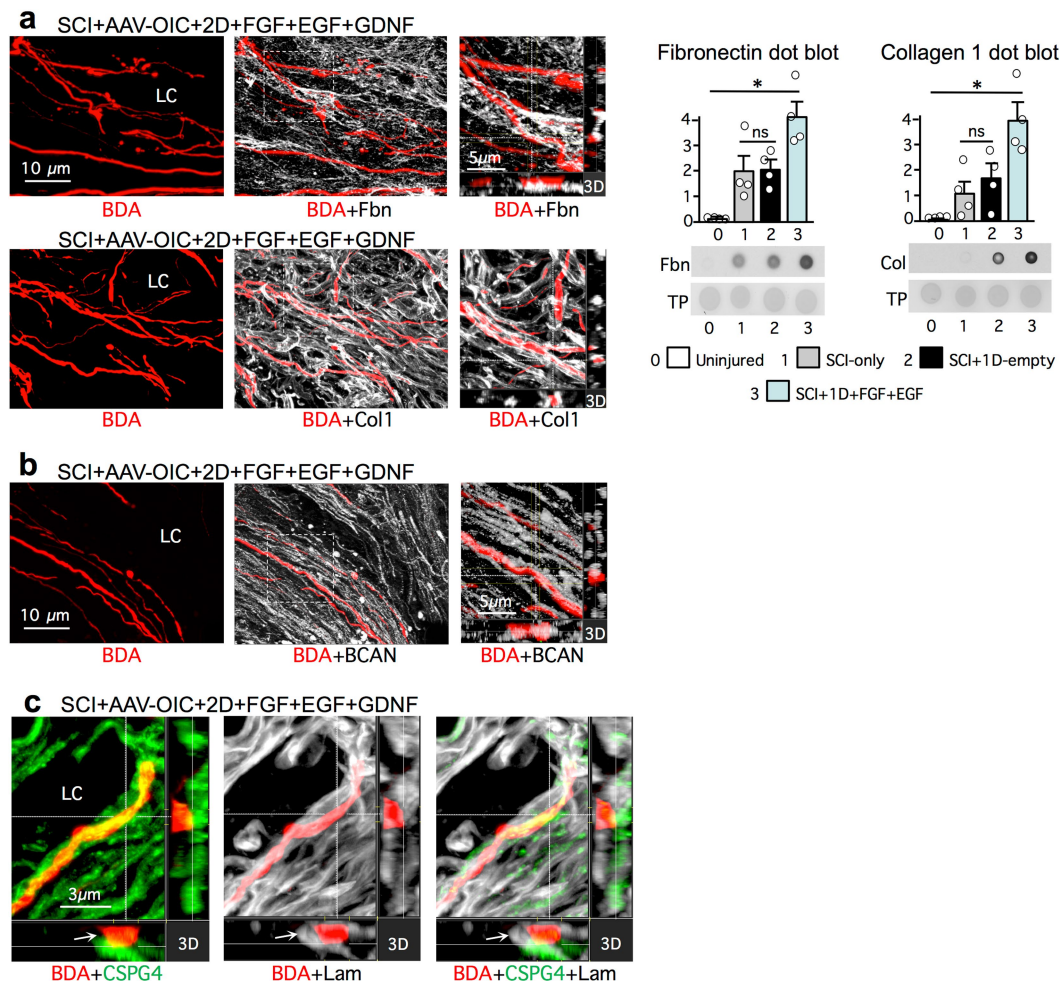
Extended Data Fig. 3 | Procedures for quantification of BDA-labelled propriospinal axons after SCI. **a**, Schematics show demarcation of SCI lesion centre (Cn) and evenly spaced lines beyond the lesion centre placed by image analysis software (NeuroLucida, MicroBrightField) for quantification of axon intercepts in horizontal tissue sections of mice with SCI and one (D₁) or two (D₁ + D₂) hydrogel depots. **b**, Multi-fluorescent survey images show BDA-labelled axons and GFAP-labelled astrocytes that demarcate astrocyte scar proximal borders and distal borders around the non-neural lesion core after SCI. The hydrogel of the empty depot (left) was tagged with a blue fluorescent label for visualization. Note the essential absence of axons passing the astrocyte scar (AS) proximal border to reach the lesion centre (Cn) or beyond in the mouse with SCI plus empty depot (left), in contrast to the large number of axons that regrow through the lesion core and passed beyond the distal astrocyte scar border into spared grey matter in the mouse with full treatment of stimulatory AAV plus growth factors (right). GFAP staining shows that the SCI lesions are anatomically complete across the entire width of the spinal cord in both cases. Note that the second depot was placed at nine days after SCI, by which time the distal astrocyte scar border was essentially formed³⁵. Note also that astrocytes do not migrate into the depots, potentially giving the mistaken impression of cavity formation when looking only at the

GFAP channel alone. Nevertheless, examination of other fluorescence channels shows that depot sites clearly contain DAPI-stained stromal cells and BDA-positive axons. **c**, Large area survey images of BDA-labelled axons in composite mosaic scans of horizontal sections. In a control mouse (top) that received SCI plus empty depot, few axons reach the lesion centre, almost none pass beyond and no axons are present at 3 mm. In the treated mouse (middle) that received stimulatory AAV plus growth factors, many axons regrow through the lesion core and reach or pass 1.5 mm beyond the lesion centre, which is the equivalent length of a full thoracic spinal segment in mice⁴⁶. Note also that there are no axons present at 3 mm, demonstrating that the SCI lesion was complete and that axons that are found past the lesion centre represent axon regrowth after SCI in response to the experimental manipulations. In an uninjured mouse (bottom), there are many labelled axons at the distance equivalent to 3 mm beyond the location of SCI in injured mice. **d**, Numbers of axon intercepts at lesion centres for all experimental groups. Data are mean ± s.e.m., dots in graphs show numbers and distribution of individual mice per group. NS, not significant versus SCI only; #P < 0.01, versus SCI only and not significant versus each other; **P < 0.01; ***P < 0.001, versus all other groups; one-way ANOVA with Bonferroni, F(12, 57) = 22.3.



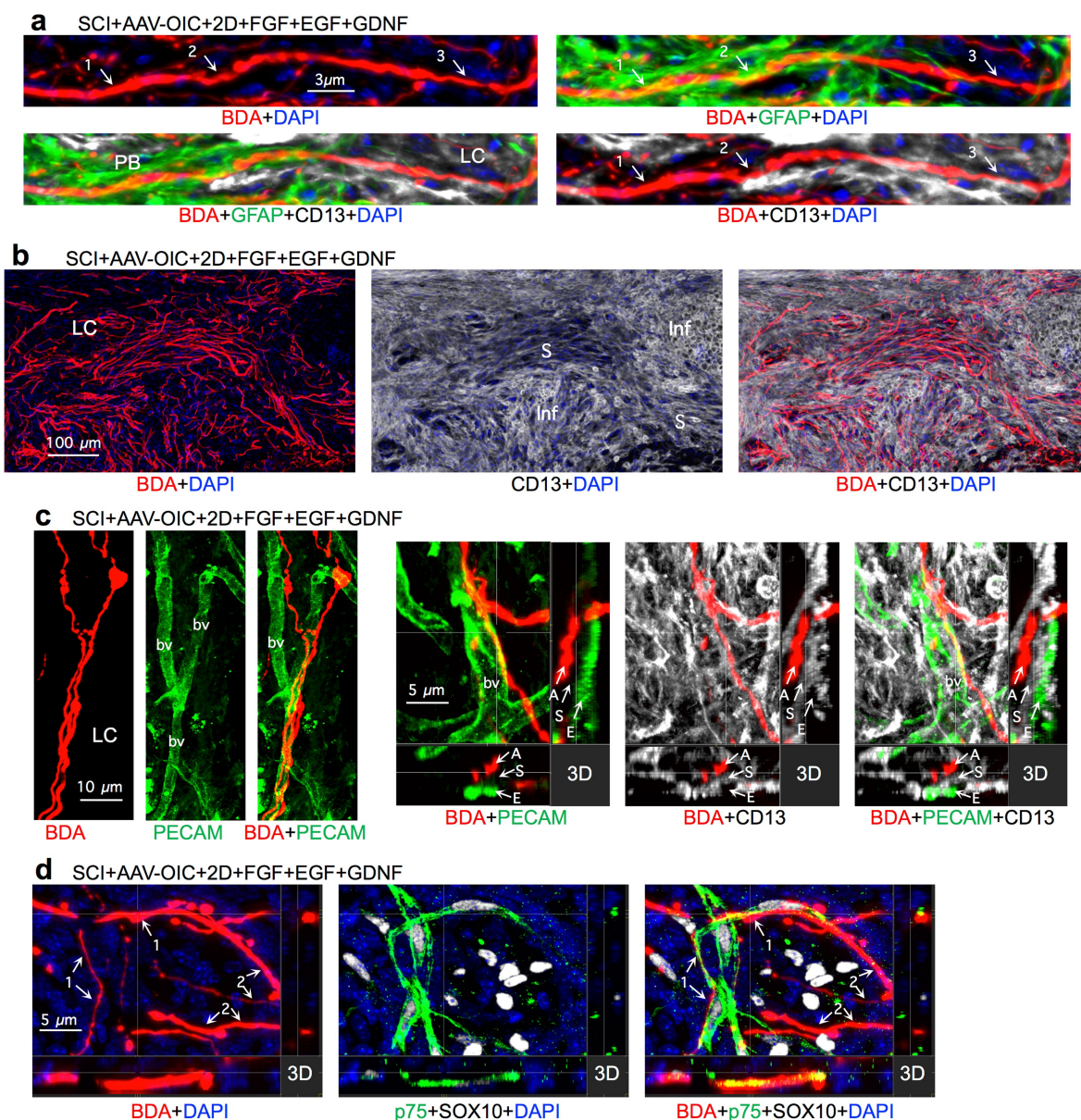
Extended Data Fig. 4 | BDA tract-tracing of propriospinal axons after SCI and different treatment conditions. Survey images show tiled mosaic scans of horizontal sections from representative mice of all experimental conditions. Experimental treatment conditions are listed in the upper left of each scan. Mouse identification numbers are given in the upper right.

Scans are oriented with their lesion centres aligned along the dashed lines so that axon growth to or past this point can be easily compared. Axon regrowth was quantified by counting axon intercepts with lines drawn through lesion centres and at regular intervals beyond by using image analysis software.



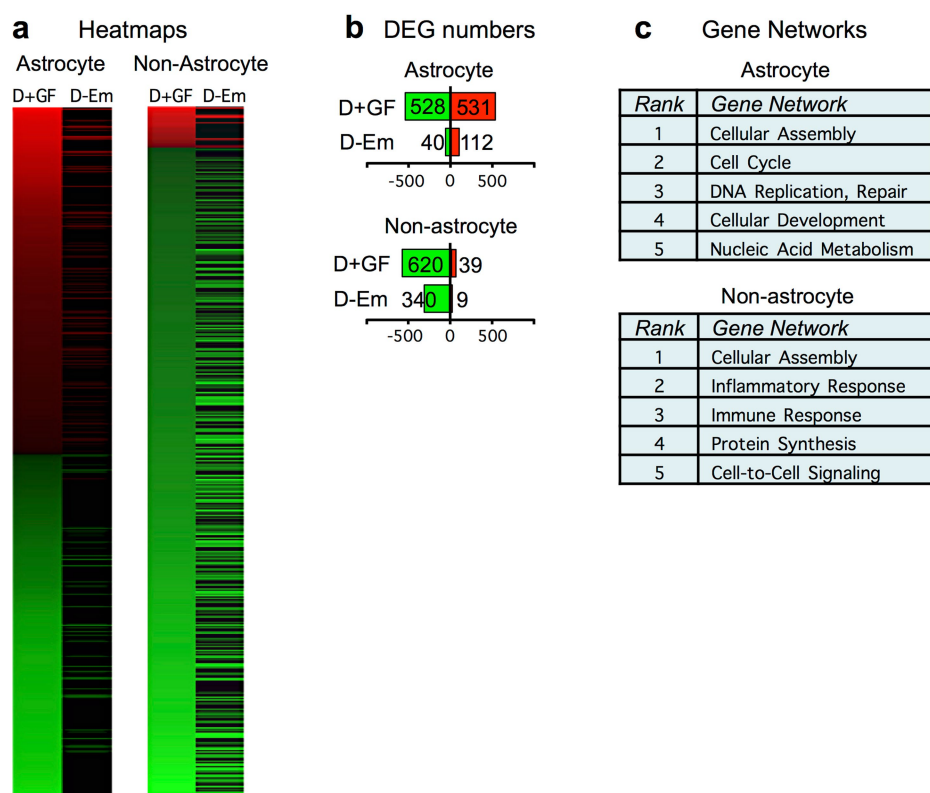
Extended Data Fig. 5 | Stimulated, supported and chemoattracted mouse propriospinal axons regrow through the lesion core in contact with various substrate molecules, including putatively inhibitory CSPGs. a. Left and left middle, multi-fluorescent, detail images show BDA-labelled axons regrowing along and among surfaces decorated with fibronectin or collagen. Right middle, orthogonal 3D confocal images of the outlined areas show direct contact between BDA-labelled axons and fibronectin or collagen. Right, quantification of fibronectin or collagen levels and dot blots. Data are mean \pm s.e.m. of density, $n = 4$ mice per

group. $*P < 0.01$, one-way ANOVA with Bonferroni, $F(3, 12) = 13.0$ for fibronectin dot blot and $F(3, 12) = 10.2$ for collagen dot blot). **b.** Left and middle, multi-fluorescent, detail images show BDA-labelled axons regrowing along and among surfaces decorated with brevican (BCAN). Right, orthogonal 3D confocal image of the outlined area shows direct contact between BDA-labelled axons and BCAN. **c.** Multi-fluorescent, orthogonal 3D confocal images show BDA-labelled axons regrowing along and in direct contact with surfaces decorated with both CSPG4 and laminin (arrows).



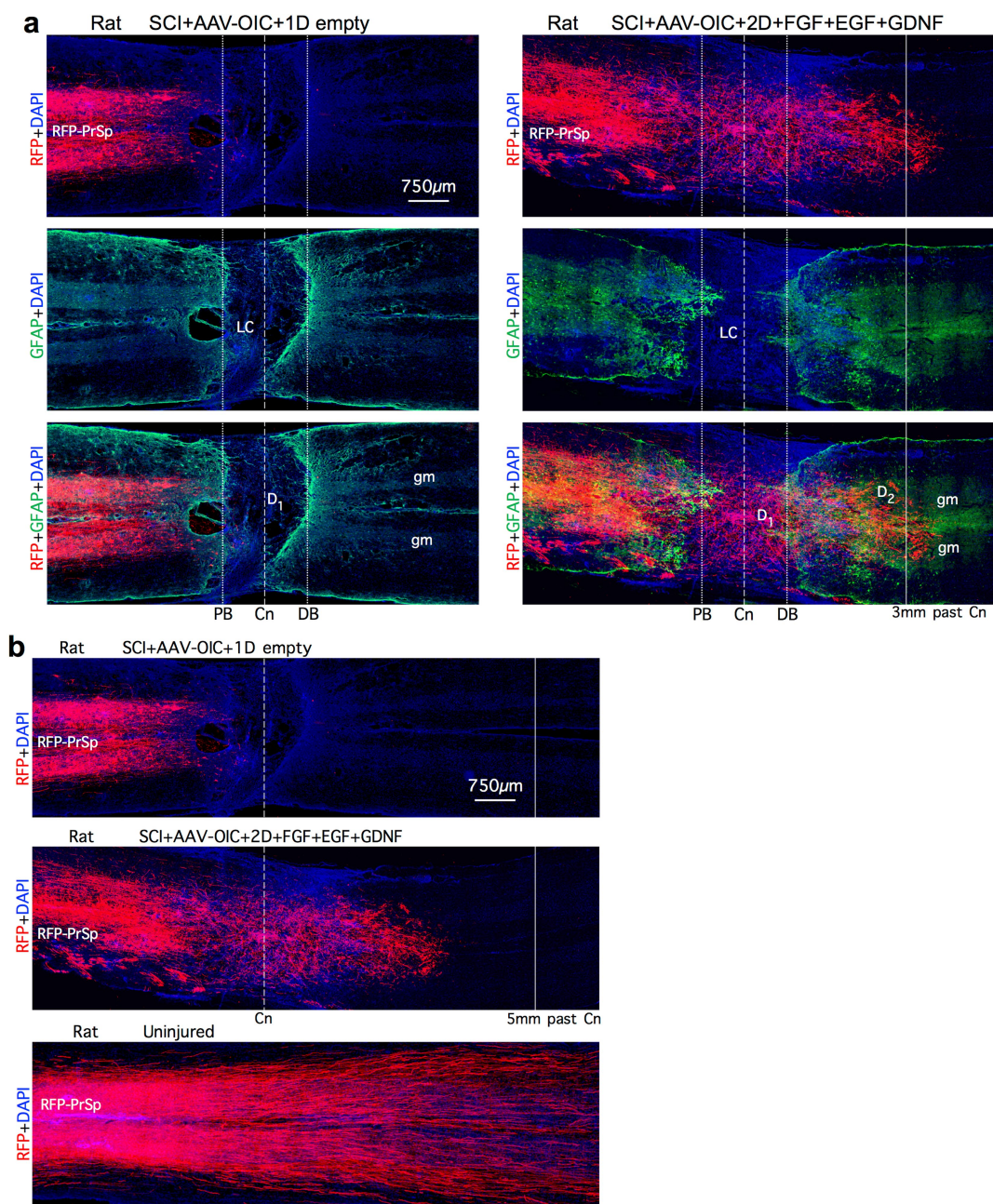
Extended Data Fig. 6 | BDA tract-tracing of propriospinal axon regrowth after SCI along and among different cell types. **a**, Multiple channel fluorescent images show the same BDA-labelled axon transitioning from contact with GFAP-positive astrocytes in proximal scar border to contact with CD13⁺ stromal cells in the lesion core. Numbers and arrows indicate the same locations in images of different combinations of fluorescent markers: 1, axons in contact with astrocyte processes; 2, axons in contact with both astrocyte process and stromal cell; 3, axons in contact with stromal cell. **b**, Multiple channel fluorescent images show axons regrowing along, and following the trajectory of, stromal cells (S) while circumventing clusters of inflammatory cells (Inf) in the lesion core. **c**, Multiple channel fluorescent images show axons (A) regrowing along

the trajectory of blood vessels (bv) in contact with stromal cells that are present on endothelia (E) positive for platelet endothelial cell adhesion molecule (PECAM). **d**, Multiple channel fluorescent images show BDA-labelled propriospinal axons and cells expressing the combinatorial Schwann cell markers, p75 and SOX10, in the lesion core. Some stromal cells expressing only SOX10 but not p75 are also visible. Numbers and arrows indicate the same locations in images of different combinations of fluorescent markers: 1, axons in partial contact with cells expressing Schwann cell markers; 2, axons not in detectable contact with Schwann cells. Note that some axons are partially in contact with, and partially not in contact with, Schwann cells in lesion core.



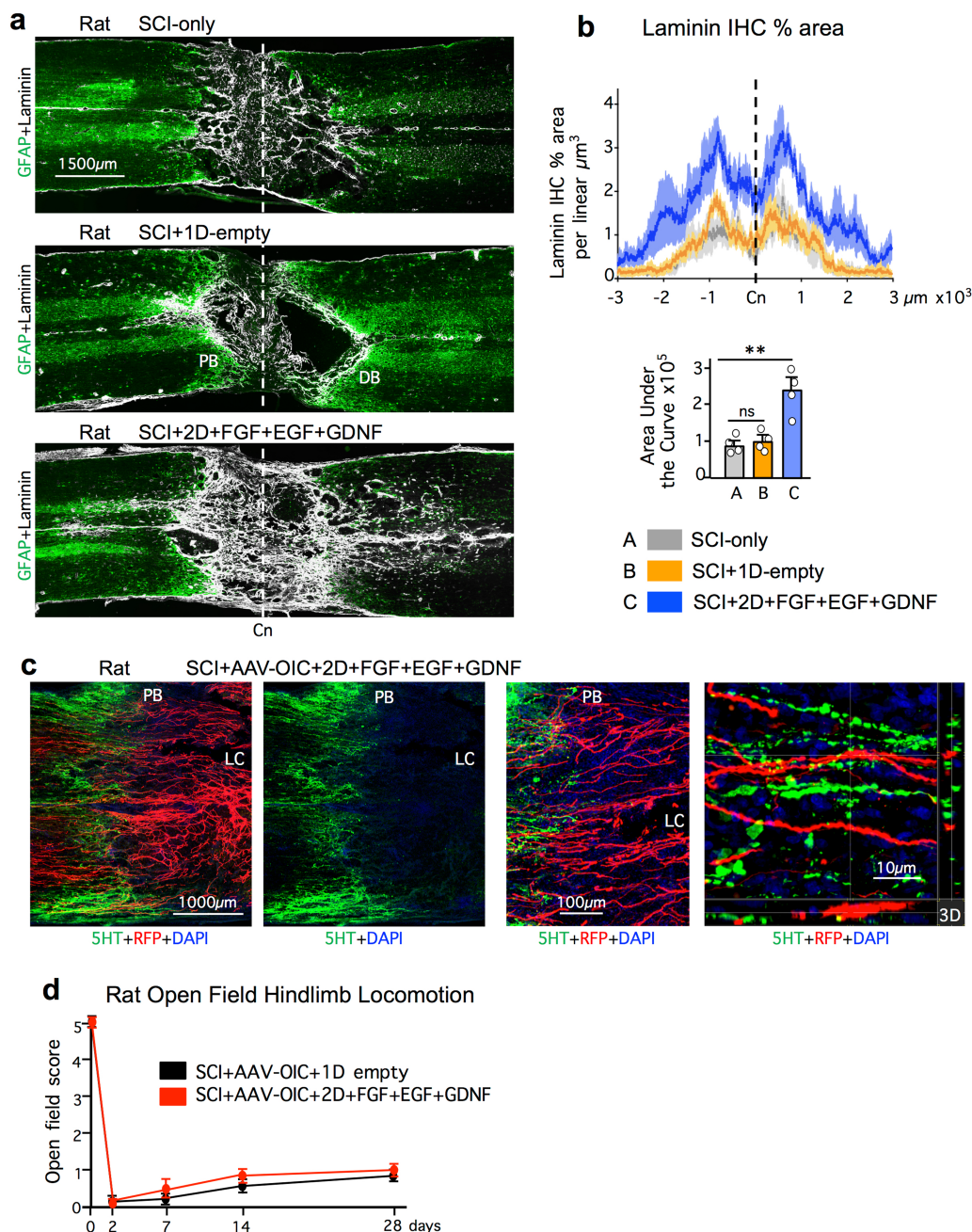
Extended Data Fig. 7 | Comparison of genomic data from astrocytes and non-astrocyte cells from mice with or without FGF + EGF after SCI. **a**, Heat maps showing significantly differentially expressed genes (DEGs) derived by RNA-seq of mRNAs from spinal cord tissue of mice treated with SCI + 1D + FGF + EGF (D + GF), and the expression of these genes in mice treated with SCI + 1D empty (D-Em), at two weeks after SCI. Data are shown for mRNAs derived selectively from astrocytes or from all other cell types (non-astrocytes), isolated as previously described¹⁹. Red, upregulated; green, downregulated; relative to SCI only. $n = 3$ mice per group; FDR < 0.1 for differential expression. **b**, Total numbers of significant DEGs in astrocytes and non-astrocytes from mice

as shown in the heat map in **a**. Red and green numerical values indicate significantly upregulated and downregulated genes, respectively. Relative to SCI only, over 900 astrocyte genes and over 300 non-astrocyte genes were significantly up- or downregulated in mice after 12 days of growth factor treatment, which were not significantly altered by treatment with empty depots. **c**, Top five networks of genes significantly altered by SCI + 1D + FGF + EGF that were not altered by SCI + 1D empty after SCI relative to SCI only, as identified by unbiased analysis (Ingenuity). Full RNA-seq data are available at NCBI Gene Expression Omnibus under accession GSE111529.



Extended Data Fig. 8 | RFP tract-tracing of propriospinal axons after SCI and different treatment conditions in rats. a, b, Large area survey images of RFP-labelled axons in composite mosaic scans of horizontal sections. Tracer-injection sites are denoted by RFP-PrSp. **a,** Multiple channel fluorescent images showing BDA-labelled axons and GFAP-labelled astrocytes that demarcate astrocyte scar proximal borders and distal borders around the non-neural lesion core after SCI. GFAP-staining shows that SCI lesions were anatomically complete across the entire width of the spinal cord, with large lesion cores in a control rat (left), and in a rat treated with stimulatory AAV plus growth factors (right). In the control

rat, few axons reach the lesion centre or beyond. In the treated rat, many axons regrow through the lesion core and reach or pass 3 mm beyond the lesion centre, which is the equivalent length of a full thoracic spinal segment in rats⁴⁷. **b,** Completeness of SCI lesions was confirmed in all rats used in qualitative and quantitative evaluations by confirming that no axons were present at 5 mm or more past lesion centres, as shown here for control rats (top) and treated rats (middle), whereas in uninjured rats, abundant labelled axons are present at an equivalent distance past the RFP-injection site.



Extended Data Fig. 9 | Growth factor induction of laminin, comparison of propriospinal and serotonin axons, and locomotor evaluations of rats after SCI without and with treatments. a, b, Survey images show laminin IHC in tiled mosaic scans of horizontal sections from representative rats. **b,** Top, mean \pm s.e.m. quantification of laminin IHC in rats as per cent area per linear μ m³, ($n = 4$ rats per group, dark coloured lines = means, lighter coloured shaded areas = s.e.m., colours indicate experimental groups as shown in graph below). Bottom, total laminin in rats summarized as mean \pm s.e.m. area under the curve as calculated from graph above. (ns non-significant, $**P < 0.005$, one-way ANOVA/Bonferroni, $F(2, 9) = 15.04$). **c,** Multiple channel fluorescent images show RFP-labelled propriospinal axons and immunohistochemically stained serotonin (5HT) axons in rats after SCI + AAV-OIC + 2D + FGF + EGF + GDNF. The

two survey images on the left show the same field with different filters. Note in the survey images on the left, and in the higher magnification image in centre, that RFP-labelled propriospinal axons regrow robustly past the astrocyte scar proximal border and through the non-neural lesion core. By contrast, 5HT axons did not regrow into or through the lesion core. The image on the right shows an orthogonal 3D confocal detail from an area proximal to the SCI lesion, demonstrating a complete lack of overlap of RFP labelling and 5HT immunohistochemistry, indicating that RFP tracing did not label 5HT axons of passage. **d,** Open field hindlimb locomotor score at various times after SCI in rats assessed using a 6-point scale in which 5 is normal walking and 0 is no movement of any kind. Data are mean \pm s.e.m., $n = 6$ per rats group.

Extended Data Table 1 | Abbreviations used in the text and figures

1D	animals receiving 1 hydrogel depot
2D	animals receiving 2 hydrogel depots
α CD29	anti-CD29 function-blocking antibody
AAV	adeno-associated virus
BCAN	brevican
BDA	biotinylated dextran amine (axon tract tracer)
CD29	integrin beta-1
Cn	lesion center
CNTF	ciliary-derived neurotrophic factor
CSPG	chondroitin sulfate proteoglycan
D ₁	hydrogel depot in lesion center
D ₂	hydrogel depot in spared neural tissue
EGF	epidermal growth factor
FGF2	fibroblast growth factor 2 (basic)
GDNF	glial derived neurotrophic factor
GFP	green fluorescent protein
IGF	insulin-like growth factor
Inf	Inflammatory cells
OIC	osteopontin, IGF plus CNTF
PECAM	platelet endothelial cell adhesion molecule
PrSp	propriospinal neurons
PTEN	phosphatase and tensin homolog
RFP	red fluorescent protein (axon tract tracer)
shPTEN	short hairpin RNA against PTEN
S	stromal cells
Syn	synaptophysin
TP	total protein

A fluid-to-solid jamming transition underlies vertebrate body axis elongation

Alessandro Mongera^{1,2,7}, Payam Rowghanian^{1,2}, Hannah J. Gustafson^{1,2,3}, Elijah Shelton^{1,2}, David A. Kealhofer⁴, Emmet K. Carn¹, Friedhelm Serwane^{1,2,8}, Adam A. Lucio^{1,2}, James Giammona^{2,4} & Otger Campàs^{1,2,5,6*}

Just as in clay moulding or glass blowing, physically sculpting biological structures requires the constituent material to locally flow like a fluid while maintaining overall mechanical integrity like a solid. Disordered soft materials, such as foams, emulsions and colloidal suspensions, switch from fluid-like to solid-like behaviours at a jamming transition^{1–4}. Similarly, cell collectives have been shown to display glassy dynamics in 2D and 3D^{5,6} and jamming in cultured epithelial monolayers^{7,8}, behaviours recently predicted theoretically^{9–11} and proposed to influence asthma pathobiology⁸ and tumour progression¹². However, little is known about whether these seemingly universal behaviours occur in vivo¹³ and, specifically, whether they play any functional part during embryonic morphogenesis. Here, by combining direct in vivo measurements of tissue mechanics with analysis of cellular dynamics, we show that during vertebrate body axis elongation, posterior tissues undergo a jamming transition from a fluid-like behaviour at the extending end, the mesodermal progenitor zone, to a solid-like behaviour in the presomitic mesoderm. We uncover an anteroposterior, N-cadherin-dependent gradient in yield stress that provides increasing mechanical integrity to the presomitic mesoderm, consistent with the tissue transitioning from a wetter to a dryer foam-like architecture. Our results show that cell-scale stresses fluctuate rapidly (within about 1 min), enabling cell rearrangements and effectively ‘melting’ the tissue at the growing end. Persistent (more than 0.5 h) stresses at supracellular scales, rather than cell-scale stresses, guide morphogenetic flows in fluid-like tissue regions. Unidirectional axis extension is sustained by the reported rigidification of the presomitic mesoderm, which mechanically supports posterior, fluid-like tissues during remodelling before their maturation. The spatiotemporal control of fluid-like and solid-like tissue states may represent a generic physical mechanism of embryonic morphogenesis.

One of the hallmarks of animal development is the formation of the anteroposterior body axis, which occurs by nearly unidirectional elongation of tissues at the anterior and posterior body ends¹⁴. During posterior elongation, which does not depend on cell proliferation at early stages¹⁵, descendants of neuromesodermal progenitors move ventrally from the dorsal medial region to the paraxial mesodermal progenitor zone (MPZ) as mesodermal progenitors^{16,17} (Fig. 1a, b). These progenitor cells progressively differentiate into mature mesodermal cells that are incorporated in the presomitic mesoderm (PSM) as body elongation proceeds. Both in amniotes and fish, cells display a graded reduction of their movements as they progress from the MPZ (referred to as posterior PSM in amniotes) to the PSM^{16,18}, where they arrest and undergo a mesenchymal-to-epithelial transition that anticipates somite formation¹⁹. The observed patterns of cell motion in these kinematic studies^{16,18} could potentially be caused by a gradient in cell or tissue mechanics along the anteroposterior axis. However, the physical mechanism causing such cellular movements and, more generally, sculpting the body axis, remains unknown.

Morphogenetic movements are generally believed to be guided by mechanical forces²⁰. To establish the role of mechanical forces in vertebrate body elongation, we measured the endogenous mechanical stresses along the anteroposterior axis of zebrafish embryos using magnetically responsive oil microdroplets²¹ (Fig. 1a–c; Methods). In the absence of an applied magnetic field, the deformations of droplets inserted between cells in the tissue provide a readout of the local mechanical stresses²². To detect potential spatial variations in mechanical stresses at supracellular scales, we analysed the ellipsoidal deformation of the droplets (Fig. 1d; Methods). Our analysis shows a posterior-to-anterior increase in the magnitude of supracellular stresses (Fig. 1e), with the axis of droplet elongation oriented mediolaterally in the medial MPZ (M-MPZ) and progressively reorienting along the anteroposterior axis in the PSM (Fig. 1f), consistent with the previously reported directions of morphogenetic flows^{16,18}. Monitoring the ellipsoidal droplet deformation over time shows that supracellular stresses persist over more than 30 min (Fig. 1g), compatible with developmental timescales (around 0.5 to 1 h). These data indicate that supracellular stresses guide morphogenetic flows and reveal an increase in mediolateral constriction during PSM maturation, consistent with mediolateral thinning of the body axis¹⁶ and anisotropy of nuclear deformation and cell shape (Extended Data Fig. 1).

Beyond supracellular stresses, the deviations in the shapes of the droplets from ellipsoidal deformations display cell-sized stress inhomogeneities (Fig. 1h, j; Extended Data Fig. 2; Methods). Both the average and maximal values of these cell-scale stresses are uniform along the anteroposterior axis (Fig. 1i). Temporal autocorrelation analysis of these higher order droplet deformations shows that cell-scale stresses are short-lived, with a persistence time of approximately 1 min (Fig. 1k; Methods). These data show that cell-scale stresses are non-persistent at developmental timescales (0.5 to 1 h) and uniform throughout the tissue, suggesting that their role may be to introduce active stress fluctuations (causing cell ‘jiggling’; Supplementary Video 1), as previously proposed for 3D multicellular aggregates²³.

Since morphogenetic flows also depend on the material properties of the tissue^{24,25} and involve large tissue rearrangements, it is important to analyse the tissue mechanical response to large deformations. To do so, we applied a controlled, uniform magnetic field for a defined time period (15 min) to droplets previously inserted in the tissue²¹ (Fig. 2a; Methods). Magnetic actuation caused large droplet deformations of more than one cell size, corresponding to applied strains in the 50–150% range (Fig. 2a, b). Upon removal of the magnetic field, droplets progressively relaxed towards a spherical shape due to the restoring force of interfacial tension (capillary stress, σ_c), but arrested before reaching this state, displaying a residual deformation and revealing the presence of a yield stress σ_y in the tissue (Fig. 2a–c; Methods). The yield stress, which corresponds to the maximal mechanical stress that a material can sustain in a solid-like state before starting to flow³, provides a direct measure of the ability of the tissue to withstand sustained mechanical

¹Department of Mechanical Engineering, University of California, Santa Barbara, CA, USA. ²California NanoSystems Institute, University of California, Santa Barbara, CA, USA. ³Biomolecular Science and Engineering Program, University of California, Santa Barbara, CA, USA. ⁴Department of Physics, University of California, Santa Barbara, CA, USA. ⁵Department of Molecular, Cell and Developmental Biology, University of California, Santa Barbara, CA, USA. ⁶Center for Bioengineering, University of California, Santa Barbara, CA, USA. ⁷Present address: European Molecular Biology Laboratory, Heidelberg, Germany. ⁸Present address: Max Planck Institute for Medical Research, Heidelberg, Germany. *e-mail: campas@ucsb.edu

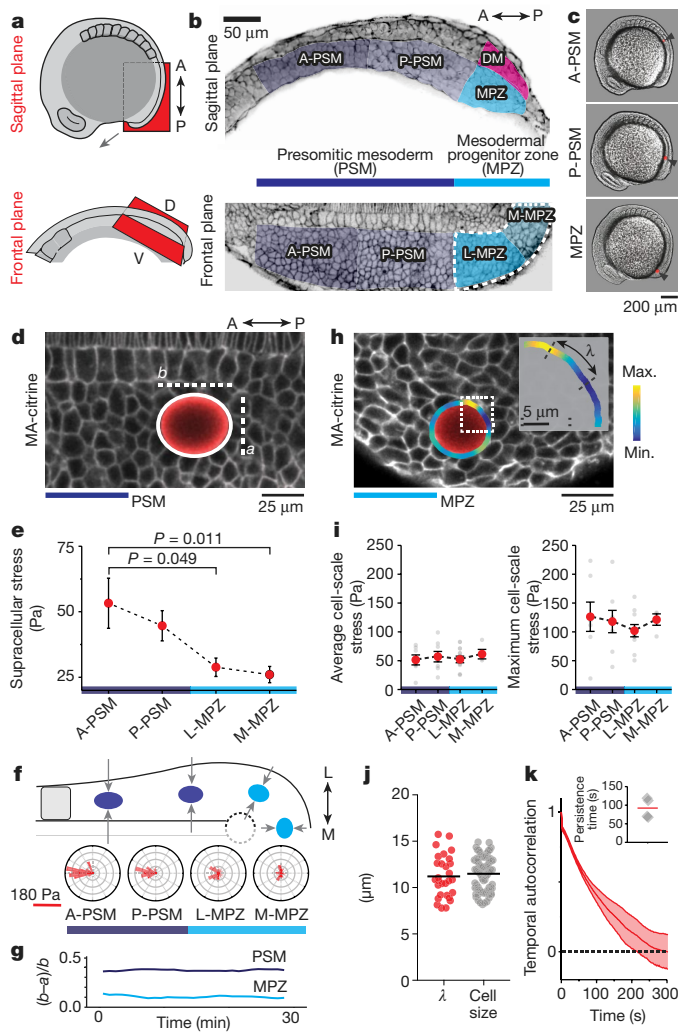


Fig. 1 | Supracellular and cell-scale mechanical stresses during body axis elongation. **a**, Sketch showing lateral views of a 10-somite stage embryo and sagittal and frontal anatomical planes. A, anterior; P, posterior; V, ventral; D, dorsal. **b**, Confocal sections along sagittal and frontal planes of posterior extending tissues in *Tg(actb2:MA-citrine)* embryos (inverted). The PSM is divided into anterior (A-PSM) and posterior (P-PSM) regions and the MPZ is divided into lateral (L-MPZ) and medial (M-MPZ) regions. The dorsal medial (DM) zone is dorsal to the MPZ. **c**, Embryos with droplets (red; arrows) located in the different regions. **d**, Elliptical fit (white; b and a being the long and short semi-axes) of a ferrofluid oil droplet (red) in the PSM of a *Tg(actb2:MA-citrine)* zebrafish embryo (no magnetic actuation). **e**, Magnitude of supracellular stresses along the anteroposterior axis ($n = 9$ (A-PSM), 24 (P-PSM), 25 (L-MPZ), 27 (M-MPZ); mean \pm s.e.m.). Mann–Whitney U -test. **f**, Bottom, orientation of the long axis of the droplets with respect to the anteroposterior axis ($n = 15$ (A-PSM), 12 (P-PSM), 11 (L-MPZ), 13 (M-MPZ)). M, medial; L, lateral. Top, sketch showing the average droplet orientations along the anteroposterior axis and the posterior-to-anterior increase in mediolateral constriction (arrows) in the PSM. **g**, Time evolution of the ellipsoidal droplet deformation, $(b - a)/b$. **h**, Ferrofluid droplet (red) in the MPZ of a *Tg(actb2:MA-citrine)* embryo. Inset, curvature values along the detected droplet contour (colour coded), with λ being the distance between consecutive curvature maxima and minima. **i**, Measured average (left) and maximal (right) cell-scale stresses along the anteroposterior axis ($n = 7$ (A-PSM), 8 (P-PSM), 10 (L-MPZ), 4 (M-MPZ); mean \pm s.e.m.). **j**, Measured values of λ ($n = 29$) and cell size ($n = 100$ cells). Line indicates mean. **k**, Temporal autocorrelation of droplet shape deviations from the ellipsoidal mode ($n = 2,062$ curvature time traces obtained from 4 embryos). Average half-life is approximately 1 min (inset; $n = 4$; line indicates mean). Unless stated otherwise, n represents number of embryos.

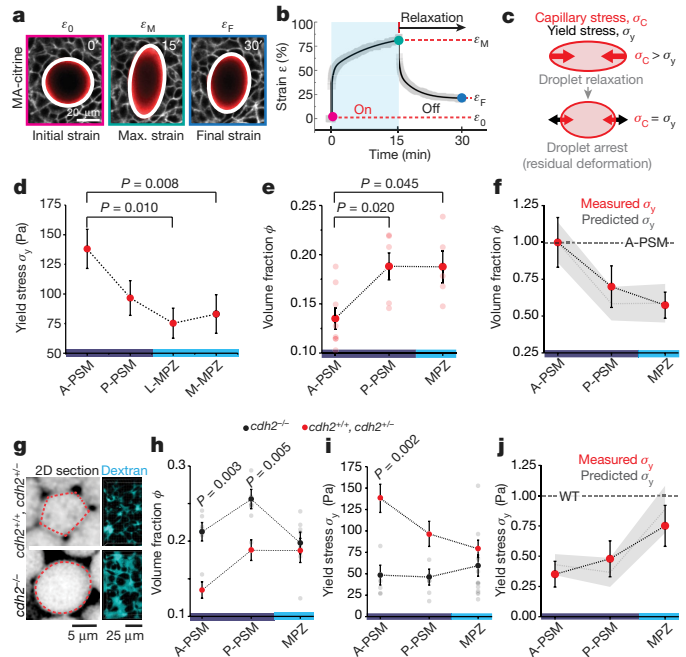


Fig. 2 | Mechanical integrity of the extending body axis. **a**, Example of droplet dynamics during magnetic actuation. White lines indicate ellipse segmentation. Droplet actuation is characterized by an initial deformation or strain (ϵ_0), a maximal strain, (ϵ_M), and a final residual strain (ϵ_F) (see Methods). **b**, Temporal evolution of the droplet strain during magnetic actuation. Experimental data points (grey) and fit (black line) are shown (Methods). **c**, The residual droplet deformation is set by the balance between the capillary stress (σ_C) and the tissue yield stress (σ_Y). **d**, **e**, Measured yield stress (**d**; $n = 12$ (A-PSM), 12 (P-PSM), 13 (L-MPZ), 13 (M-MPZ)) and volume fraction of extracellular space (ϕ) (**e**; $n = 8$ (A-PSM), 6 (P-PSM), 5 (MPZ)) along the anteroposterior axis. Mean \pm s.e.m.; Mann–Whitney U -test. **f**, Comparison between measured (red dots) and predicted (grey line and band representing mean \pm s.e.m.) yield stress along the anteroposterior axis, relative to the A-PSM. **g**, Confocal sections (inverted) and 3D reconstructions (Methods) of dextran-labelled extracellular space for control ($cdh2^{+/+}$, $cdh2^{+/-}$) and $cdh2^{-/-}$ embryos. **h**, **i**, Measured volume fraction of extracellular space (**h**; $n = 6$ (A-PSM), 8 (P-PSM), 7 (MPZ)) and yield stress (**i**; $n = 5$ (A-PSM), 5 (P-PSM), 12 (MPZ)) along the anteroposterior axis in $cdh2^{-/-}$ embryos compared to control embryos. Mean \pm s.e.m.; Mann–Whitney U -test. **j**, Measured (red dots) and predicted (grey line and band representing mean \pm s.e.m.) yield stress values in $cdh2^{-/-}$ embryos normalized to values for control embryos in each region. In all cases, n represents number of embryos.

loads, that is, its mechanical integrity. Performing these experiments along the body axis shows that the yield stress is spatially graded and minimal in the MPZ (Fig. 2d; Methods), revealing a posterior-to-anterior gradient of increasing tissue mechanical integrity that parallels PSM maturation.

The disordered cellular structure of tissues strongly resembles that of aqueous foams^{9,23,26}, which display a finite yield stress when the volume fraction ϕ of fluid between bubbles becomes lower than a critical value ϕ_c at the jamming transition^{1,2} (Supplementary Fig. 1). Following the analogy with foams, we explored the potential role of extracellular spaces in controlling tissue yield stress. We measured the volume fraction of extracellular spaces along the anteroposterior axis using fluorescent dextran (Methods) and observed decreasing spaces between cells away from the posterior end (Fig. 2e). Since cohesion between cells depends on their adhesion strength, and N-cadherin (also known as cadherin 2, encoded by the *cdh2* gene) has been shown to be essential for body axis elongation^{16,27} and PSM maturation²⁸ in zebrafish, we measured the volume fraction of extracellular spaces and yield stress along the anteroposterior axis of *cdh2* loss-of-function mutants²⁷. Disruption of N-cadherin-mediated adhesion resulted in larger extracellular spaces in the PSM, but not in the MPZ (Fig. 2g, h;

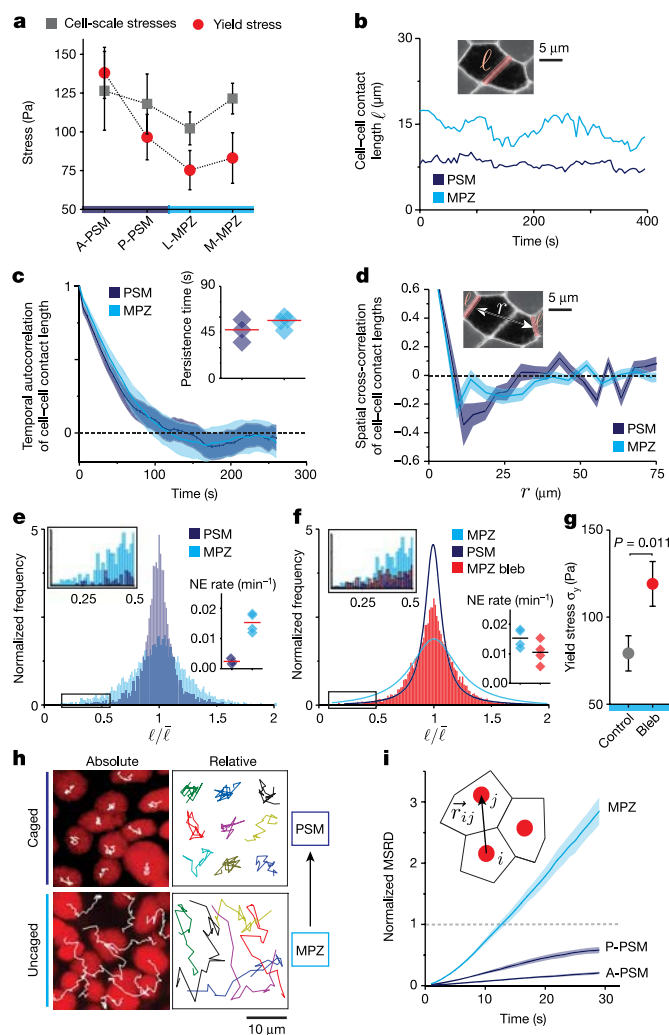


Fig. 3 | Cell-cell contact length fluctuations and cellular movements along the anteroposterior axis. **a**, Comparison of yield stress (Fig. 2d) and maximal cell-scale stresses (Fig. 1i) along the anteroposterior axis. **b**, Examples of time traces of cell-cell contact length ℓ in the PSM and MPZ. **c**, **d**, Temporal autocorrelation of cell-cell contact length (**c**) and spatial cross-correlation of cell-cell contact lengths separated by a distance r (**d**) in the PSM and MPZ ($n = 186$ (PSM), 54 (MPZ) cell-cell contacts). Correlation (persistence) timescale (**c**, inset; line indicates mean). **e**, Normalized frequency of cell-cell contact lengths in the PSM and MPZ ($n = 6,969$ (PSM), 7,896 (MPZ) cell-cell contacts). Zoomed in distribution tail (box, left inset). Neighbour exchange (NE) rates (per cell) in PSM and MPZ (right inset; $n = 14$ (PSM), 49 (MPZ) neighbour exchanges; line indicates mean). In **c–e**, cell-cell contacts were obtained from 3 and 4 embryos in PSM and MPZ, respectively. **f**, Normalized frequency of cell-cell contact lengths in the MPZ of embryos treated with 100 μM blebbistatin (bleb) ($n = 13,813$ cell-cell contacts from 4 embryos). Zoomed in distribution tail (box, left inset). Neighbour exchange rate in the MPZ region of blebbistatin-treated embryos compared to untreated embryos (right inset; $n = 20$ neighbour exchanges from 3 embryos). **g**, Measured yield stress in the MPZ of embryos treated with blebbistatin (red; $n = 12$ embryos) compared to untreated embryos (grey, $n = 26$ embryos). Mean \pm s.e.m.; Mann-Whitney U -test. **h**, **i**, Tracks (**h**; absolute and relative; nuclear tracking) and normalized mean square relative displacement (MSRD) (**i**; $n = 1,937$ (A-PSM), 1,776 (P-PSM) and 2,523 (MPZ) analysed cell pairs, from 5, 5 and 6 embryos, respectively) of cellular movements during a 30-min time window.

Extended Data Fig. 3). Accordingly, the yield stress in *cdh2* mutants is altered in the PSM but not in the MPZ (Fig. 2i). No significant spatial variation in yield stress was observed in *cdh2* mutants, indicating that N-cadherin-mediated adhesion is necessary to maintain the anteroposterior gradient in extracellular spaces and tissue mechanical integrity

(yield stress). In addition, the reported anteroposterior gradients in supracellular stresses (Fig. 1e) and nuclear anisotropy observed in control embryos are lost in *cdh2*^{−/−} mutants (Extended Data Fig. 1), indicating that mediolateral constriction in the PSM is N-cadherin-dependent. Further supporting the analogy with foams, both the measured values of yield stress along the anteroposterior axis in wild-type embryos and the changes in yield stress between control and *cdh2*^{−/−} mutant embryos can be accounted for (with no adjustable parameters) by assuming that such changes result solely from the observed variations in extracellular volume fraction in a tissue that behaves like a 3D-disordered monodispersed foam (Fig. 2f, j; Methods). Cell rearrangements induced by droplet actuation (Extended Data Fig. 4) further indicate that the yield stress arises from the jammed cellular environment, as in 3D disordered foams¹. Whereas jamming transitions have recently been predicted by 2D vertex models of multicellular systems with no extracellular spaces^{10,11} and observed in epithelial cell culture⁸, our measurements indicate that jamming in 3D embryonic tissues is consistent with classical jamming scenarios in foams, where the volume fraction of extracellular spaces has a prominent role^{1,2}.

Unlike aqueous foams, however, living tissues are characterized by active cellular stresses that could locally unjam the system and fluidize the tissue, as observed in systems for which thermal fluctuations are relevant^{2,4,9}. Comparison of the yield stress and the magnitude of cell-scale stresses shows that the maximal stress fluctuations are well above the yield stress in the MPZ, but below it in the anterior PSM (A-PSM, Fig. 3a), indicating that actively generated stress fluctuations may help to fluidize the tissue at its posterior end. Since neighbour exchanges are necessary to fluidize foam-like tissues^{1,9} and actomyosin-driven fluctuations in cell-cell contact length cause cellular rearrangements²⁰, we characterized the dynamics of cell-cell contacts both in the PSM and MPZ (Fig. 3b–f; Methods). Autocorrelation analysis shows that cell-cell contact lengths become temporally uncorrelated within less than 1 min and spatially uncorrelated beyond 2–3 cell sizes (Fig. 3c, d), consistent with our measurements of cell-scale stresses (Fig. 1i, k) and indicating that, at developmental timescales (approximately 0.5–1 h), cell-cell contact dynamics can be seen as short-lived, active fluctuations. As only large fluctuations can induce neighbour exchanges and fluidize the tissue, we measured the distribution of cell-cell contact lengths both in the PSM and MPZ tissues (Methods). Our results show a broader distribution in the MPZ, with large fluctuations being considerably more frequent in this region than in the PSM (Fig. 3e), both in control and in *cdh2* mutant embryos (Extended Data Fig. 5). Accordingly, we found that the neighbour exchange rate is 6 times higher in the MPZ than in the PSM (Fig. 3e, inset), consistent with the MPZ being in a fluid-like state and the PSM in a solid-like state. No systematic alignment of neighbour exchanges along a single spatial direction was observed (Extended Data Fig. 6), indicating that cell intercalation in these tissues does not contribute to posterior elongation. Interpreting active fluctuations as an effective temperature^{9,18,23}, a key parameter in the control of jamming transitions^{2,4}, it is possible to obtain the energy landscape of neighbour exchanges in both MPZ and PSM (Extended Data Fig. 7; Methods). In this framework, reducing fluctuations in cell-cell contact length (lower effective temperature) by impairing myosin II-dependent force generation at the cell cortex should render the MPZ more solid-like. Partial inhibition of myosin activity with blebbistatin led to smaller cell-cell contact length fluctuations (narrower distribution; Fig. 3f), decreased neighbour exchange rate (Fig. 3f, inset) and increased yield stress in the MPZ (Fig. 3g), thereby rigidifying the tissue and leading to a reduction in both cell movements and body elongation speed (Extended Data Fig. 8). These results indicate that while the PSM is jammed in a solid-like state, with weak cell-cell contact length fluctuations (low effective temperature) unable to fluidize the tissue, the stronger cell-cell contact length fluctuations (high effective temperature) in the MPZ cause substantial cell rearrangements that effectively ‘melt’ the tissue into a fluid-like state.

Since cellular movements result from the combined effect of active stresses and mechanical constraints (tissue material properties), their

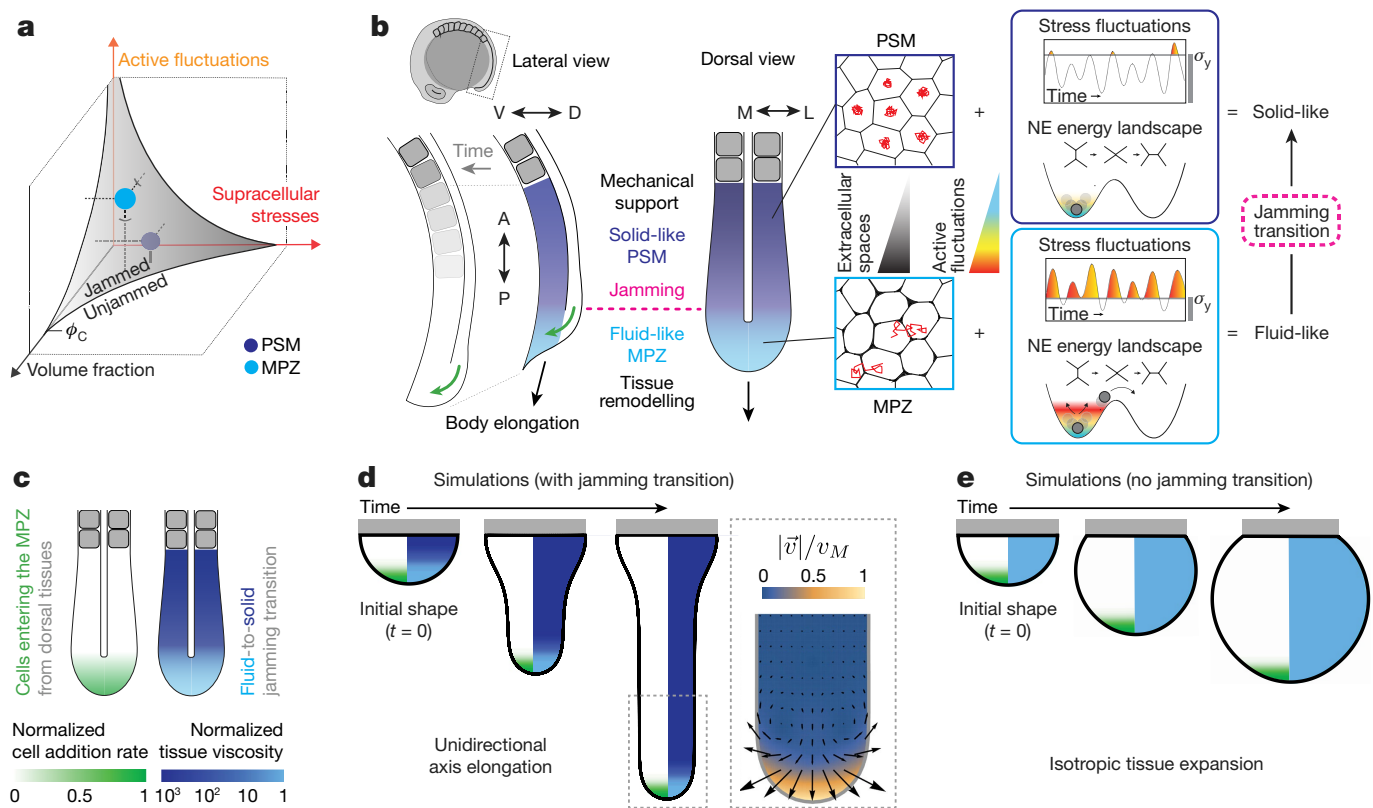


Fig. 4 | Physical mechanism of vertebrate body axis elongation.

a, The MPZ (fluid-like) and PSM (solid-like) tissue states are represented in the jamming phase diagram²⁴, which organizes jammed (solid-like) and unjammed (fluid-like) phases for varying volume fraction of extracellular spaces, supracellular stresses and active fluctuations (effective temperature). **b**, The higher cell–cell contact length fluctuations (high effective temperature) in the less constrained environment (more extracellular spaces) of the MPZ (light blue) drive cell rearrangements and cell mixing, effectively ‘melting’ the tissue in this region. As the paraxial mesoderm matures, the smaller extracellular spaces and low cell–cell contact fluctuations (low effective temperature) in the PSM (violet) rigidify the tissue via a jamming transition. Cells entering the MPZ from the dorsal medial region (green arrows; lateral view) cause the expansion of the fluid-like MPZ tissue, with the solid-like PSM acting as a rigid support that biases tissue expansion towards the posterior direction,

thereby elongating the body axis. Whereas persistent mediolateral supracellular stresses restrict lateral tissue expansion, the main role of non-persistent cell-scale stresses is to ‘melt’ the MPZ tissue, enabling its expansion and posterior elongation upon addition of new cells from the dorsal medial region. **c**, Sketch of the posterior body showing the input physical fields in the simulation. **d**, **e**, Time evolution of simulated tissue shapes (black outline) in the presence (**d**) and absence (**e**) of a jamming transition along the anteroposterior axis. The colour code in the right half of each shape corresponds to the spatial profile of the tissue viscosity (diverging in solid-like tissue regions), whereas the left half shows the anteroposterior profile of cell ingress rate into the MPZ from dorsal medial tissues. Grey rectangles represent a rigid boundary. Simulated morphogenetic flows (velocity field: direction, arrows; magnitude, colour coded) in the presence of a jamming transition leading to unidirectional body elongation (**d**, inset).

analysis informs about the mechanical state of the tissue^{6,8}. Both direct observation (Fig. 3h) and analysis (Fig. 3i; Methods) of cellular movements in the tissue showed caged behaviour in the PSM, consistent with cells being trapped in a solid-like material, and uncaged cell mixing in the MPZ, consistent with this region being fluid-like. Our measurements indicate that the posterior-to-anterior transition between uncaged to caged cellular movements, consistent with previous observations in chicken¹⁸, results from an increase in physical constraints on cells, rather than a decrease in cellular stresses, as the spaces between cells are reduced anteriorly while maintaining a constant cell density (Extended Data Fig. 9; Methods).

Taken together, our direct *in vivo* measurements of tissue mechanics and analysis of cell rearrangements and movements are all consistent with the tissue behaving as a disordered, glassy material undergoing a jamming transition as the MPZ progressively rigidifies into the PSM (Fig. 4a, b). The solid-like state of the PSM helps to maintain tissue architecture and mechanically supports the extending, fluid-like end of the posterior body, as cells from the dorsal medial region are added to the MPZ and progressively elongate the body axis (Fig. 4b). Indeed, simulations of tissue morphogenesis based solely on first principles (Fig. 4c–e; Methods; Supplementary Note 1) show that unidirectional body axis elongation naturally occurs in the presence of the observed fluid-to-solid jamming transition (Fig. 4d; Supplementary Video 2).

In absence of jamming transition, the tissue expands isotropically like a growing spherical blob and no unidirectional axis elongation occurs (Fig. 4e; Supplementary Video 3). The predicted tissue morphogenetic flows in the presence of the reported jamming transition display high posterior-directed velocities at the posterior end, no tissue flow in the A-PSM, and the existence of two counter-rotating vortices as the tissue transits from fluid-like to solid-like states (Fig. 4d, inset). Remarkably, all these predicted features in tissue flows have been observed experimentally¹⁶, indicating that the fluid-to-solid jamming transition is essential for proper unidirectional axis elongation. Phenotypes such as curved tails and the reduced elongation speed in *cdh2* mutants (Extended Data Fig. 8) can be explained by the reported loss of mechanical integrity in the PSM (Fig. 2i) during body axis elongation.

More generally, the spatiotemporal control of fluid-like and solid-like tissue states, enabling or restricting morphogenetic flows, represents a novel mechanism of tissue morphogenesis. The previously observed oscillations in actomyosin contractility during apical constriction²⁹ in *Drosophila* epithelia may be necessary to overcome a yield stress and fluidize an otherwise jammed tissue¹³. Beyond cellular jamming, transitions between fluid-like and solid-like states in tissues containing extracellular matrix between cells are likely to occur via control of the physical-chemical state of the extracellular matrix. As suggested by D’Arcy Thompson a century ago²⁶, the shaping of living tissues shares

fundamental physical principles with the sculpting of many inert materials, in which fluid-to-solid transitions are necessary to mould them into functional shapes.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0479-2>.

Received: 8 September 2017; Accepted: 3 August 2018;

Published online 5 September 2018.

1. Cohen-Addad, S., Höhler, R. & Pitois, O. Flow in foams and flowing foams. *Annu. Rev. Fluid Mech.* **45**, 241–267 (2013).
2. Liu, A. J. & Nagel, S. R. Jamming is just not cool any more. *Nature* **396**, 21–22 (1998).
3. Bonn, D., Denn, M. M., Berthier, L., Divoux, T. & Manneville, S. Yield stress materials in soft condensed matter. *Rev. Mod. Phys.* **89**, 035005 (2017).
4. Trappe, V., Prasad, V., Cipelletti, L., Segre, P. N. & Weitz, D. A. Jamming phase diagram for attractive particles. *Nature* **411**, 772–775 (2001).
5. Angelini, T. E. et al. Glass-like dynamics of collective cell migration. *Proc. Natl Acad. Sci. USA* **108**, 4714–4719 (2011).
6. Schotz, E. M., Lanio, M., Talbot, J. A. & Manning, M. L. Glassy dynamics in three-dimensional embryonic tissues. *J. R. Soc. Interface* **10**, 20130726 (2013).
7. Sadati, M., Taheri Qazvini, N. T., Krishnan, R., Park, C. Y. & Fredberg, J. J. Collective migration and cell jamming. *Differentiation* **86**, 121–125 (2013).
8. Park, J.-A. et al. Unjamming and cell shape in the asthmatic airway epithelium. *Nat. Mater.* **14**, 1040–1048 (2015).
9. Bi, D., Lopez, J. H., Schwarz, J. M. & Manning, M. L. Energy barriers and cell migration in densely packed tissues. *Soft Matter* **10**, 1885–1890 (2014).
10. Bi, D., Lopez, J. H., Schwarz, J. M. & Manning, M. L. A density-independent rigidity transition in biological tissues. *Nat. Phys.* **11**, 1074–1079 (2015).
11. Farhadifar, R., Röper, J.-C., Aigouy, B., Eaton, S. & Jülicher, F. The influence of cell mechanics, cell–cell interactions, and proliferation on epithelial packing. *Curr. Biol.* **17**, 2095–2104 (2007).
12. Oswald, L., Grosser, S., Smith, D. M. & Käs, J. A. Jamming transitions in cancer. *J. Phys. D Appl. Phys.* **50**, 483001 (2017).
13. Atia, L. et al. Geometric constraints during epithelial jamming. *Nat. Phys.* **14**, 613–620 (2018).
14. Bénazéraf, B. & Pourquié, O. Formation and segmentation of the vertebrate body axis. *Annu. Rev. Cell Dev. Biol.* **29**, 1–26 (2013).
15. Zhang, L., Kendrick, C., Julich, D. & Holley, S. A. Cell cycle progression is required for zebrafish somite morphogenesis but not segmentation clock function. *Development* **135**, 2065–2070 (2008).
16. Lawton, A. K. et al. Regulated tissue fluidity steers zebrafish body elongation. *Development* **140**, 573–582 (2013).
17. Kimelman, D. Tales of tails (and trunks): forming the posterior body in vertebrate embryos. *Curr. Top. Dev. Biol.* **116**, 517–536 (2016).
18. Bénazéraf, B. et al. A random cell motility gradient downstream of FGF controls elongation of an amniote embryo. *Nature* **466**, 248–252 (2010).
19. Aulehla, A. & Pourquié, O. Signaling gradients during paraxial mesoderm development. *Cold Spring Harb. Perspect. Biol.* **2**, a000869 (2010).
20. Heisenberg, C.-P. & Bellaiche, Y. Forces in tissue morphogenesis and patterning. *Cell* **153**, 948–962 (2013).
21. Serwane, F. et al. In vivo quantification of spatially varying mechanical properties in developing tissues. *Nat. Methods* **14**, 181–186 (2017).
22. Campàs, O. et al. Quantifying cell-generated mechanical forces within living embryonic tissues. *Nat. Methods* **11**, 183–189 (2014).
23. Marmottant, P. et al. The role of fluctuations and stress on the effective viscosity of cell aggregates. *Proc. Natl Acad. Sci. USA* **106**, 17271–17275 (2009).
24. Campàs, O. A toolbox to explore the mechanics of living embryonic tissues. *Semin. Cell Dev. Biol.* **55**, 119–130 (2016).
25. Miller, C. J. & Davidson, L. A. The interplay between cell signalling and mechanics in developmental processes. *Nat. Rev. Genet.* **14**, 733–744 (2013).
26. Thompson, D. W. *On Growth and Form* (Cambridge Univ. Press, Cambridge, 1917).
27. Lele, Z. et al. *parachute/n-cadherin* is required for morphogenesis and maintained integrity of the zebrafish neural tube. *Development* **129**, 3281–3294 (2002).
28. Chal, J., Guillot, C. & Pourquié, O. PAPC couples the segmentation clock to somite morphogenesis by regulating N-cadherin-dependent adhesion. *Development* **144**, 664–676 (2017).
29. Martin, A. C., Kaschube, M. & Wieschaus, E. F. Pulsed contractions of an actin–myosin network drive apical constriction. *Nature* **457**, 495–501 (2009).

Acknowledgements We thank E. Sletten for sharing custom-made fluorinated dyes. We also thank all laboratory members and the UCSB Animal Research Center for support. P.R. thanks B. Aigouy for assistance with Tissue Analyzer. A.M. thanks EMBO (EMBO ALTF 509-2013), Errett Fisher Foundation and Otis Williams Fund for financial support. This work was partially supported by the National Science Foundation (CMMI-1562910) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (R21HD084285; R01HD095797).

Reviewer information *Nature* thanks J. Fredberg, P.-F. Lenne, O. Pourquié and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.M. and O.C. designed research; A.M., H.J.G., D.A.K. and F.S. performed experiments; A.M., H.J.G., P.R., E.S. and J.G. analysed the data; P.R. and O.C. performed theoretical interpretation of experiments; E.K.C. and P.R. performed simulations; A.A.L. assisted with droplet generation; A.M. and O.C. wrote the paper; O.C. supervised the project.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0479-2>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0479-2>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to O.C.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Zebrafish husbandry, fish lines, and experimental manipulations. Zebrafish (*Danio rerio*) were maintained as previously described³⁰. Animals were raised and experiments were performed following all ethical regulations and according to protocols approved by the Institutional Animal Care and Use Committee (IACUC) at the University of California, Santa Barbara. For ubiquitous labelling of cell membranes, we used either Tg(*actb2*:MA-citrine) embryos^{27,31} or embryos that had been injected with membrane-GFP mRNA at the one-cell stage. For experiments to perturb the amount of extracellular spaces we used *cdh2*^{tm101/tm101} embryos (also known as parachute (pac) mutants²⁷), which feature a truncated N-cadherin (Cdh2) extracellular domain and have been shown to display cell–cell adhesion defects²⁷. For tracking of cell movements, one-cell stage embryos were injected with H2B-RFP mRNA.

Generation and injection of ferrofluid droplets. The composition and preparation of ferrofluid droplets was identical to those previously described²¹. In brief, we prepared biocompatible fluorocarbon-based ferrofluids by diluting DFF1 (Ferrotec) in Novec 7300 (3M) at varying concentrations to achieve the necessary magnetic stresses to generate the desired droplet deformations (applied strains). To prevent non-specific adhesion between the cells and the ferrofluid droplets, a fluorinated Krytox-PEG(600) surfactant (008-fluorosurfactant, RAN Biotechnologies³²) was diluted in the ferrofluid at a 2.5% (w/w) concentration. The ferrofluid used in each experiment was calibrated as previously described²¹, so that the stresses applied by the ferrofluid droplet on the surrounding material in the presence of a uniform magnetic field were quantitatively known. Ferrofluid droplets were generated inside embryos by direct injection of the ferrofluid in the tissue of interest, as previously described²¹. Control of droplet size was achieved by controlling the injection pressure and time of the injection pulse. Droplets were injected in the MPZ tissue at the 4- and 6-somite stages for measurements in the PSM and MPZ, respectively. Droplets were imaged at least 2 h after injection to allow enough time for the tissue to recover from the injection. The mechanical relaxation timescales in the MPZ and PSM are short (~1 min; Supplementary Fig. 2b) and cell rearrangements are observed in timescales shorter than 10 min (Fig. 3e, inset). Therefore, our experiments, performed 2 h after injection, allow enough time for the tissue to remodel. Indeed, no trace of the injection is observed in the tissue when imaged. Importantly, we have previously shown that injection and actuation of magnetic droplets located in the MPZ and PSM tissues do not affect normal developmental processes and, especially, body axis elongation²¹.

Imaging. Embryos were mounted for imaging in 0.8% low-melting agarose and imaged at 25 °C using a laser scanning confocal (LSM 710, Carl Zeiss Inc.). Images were taken at 2.5-s intervals for droplet actuation experiments and at 60-s intervals for analysis of cell movements, using a 40× water immersion objective (LD C-Apochromat 1.1 W, Carl Zeiss) or a 10× air objective (EC Epiplan-Neofluar 0.25, Carl Zeiss). Imaging of ferrofluid droplets in the embryo was done as previously reported²¹. Ferrofluid droplets were fluorescently labelled using a custom-synthesized fluorinated rhodamine dye³³, which was dissolved in the fluorocarbon-based ferrofluid oil at a final concentration of 37 μM.

Measurement of supracellular (tissue-level) stresses. To quantify supracellular stresses, we measured the droplet ellipsoidal deformation, as this deformation mode reveals mechanical stresses that occur at the length scale of the droplet diameter (and slightly larger scales), which we purposely made larger than the cell size (Fig. 1d; droplet diameter: 44 ± 6 μm; cell diameter: 11 ± 5 μm; mean ± s.e.m.). We used ferrofluid droplets because their interfacial tension can be calibrated in vivo²¹. No magnetic field was applied on the droplet before or during the measurement of supracellular stresses. Moreover, we prevented cell adhesion to the droplets to reveal the stresses associated solely with (supracellular) morphogenetic flows, mirroring previous studies in inert fluids³⁴. Two hours after injection, the droplet mid-plane was imaged using confocal microscopy (Zeiss LSM 710, Zeiss). To quantify the ellipsoidal deformation of the droplet (and neglect high order deformations), we fitted an ellipse to the measured droplet shape, as previously described²¹. The value of anisotropic stresses for the ellipsoidal droplet deformation is given by $\sigma_T^A = 2\gamma(H_b - H_a)$, where H_b and H_a are the mean curvatures of the droplet at the intersection of the two principal axis with the ellipsoid and γ is the droplet's interfacial tension (Fig. 1d), as previously established^{22,35}. The mean curvatures H_b and H_a are given by $H_b = b/a^2$ and $H_a = 1/2a + a/(2b^2)$, with b and a being the long and short semi-axis of the fitted ellipse, respectively (Fig. 1d). At the end of each experiment, that is, after recording the shape of each droplet, we measured its interfacial tension γ in situ, within the developing embryo, as previously described²¹.

To obtain the direction of droplet deformation, we measured the angle between the direction defined by the long axis of elliptical droplet deformation and the anteroposterior axis of the embryo using ImageJ.

While droplets cannot measure isotropic stresses because of droplet incompressibility²², the ellipsoidal droplet deformation provides a quantitative measure of the difference in stresses along the direction of droplet elongation and the

perpendicular direction in the observation plane (the principal directions of the deformation). Ellipsoidal droplet deformations can, in general, be caused both by shear stresses or by stress differences along principal axes of droplet deformation (stress anisotropy).

Droplet shape segmentation and measurement of in-plane curvature. Confocal sections through the middle of a ferrofluid droplet were obtained by confocal microscopy, as described above. We then used a custom-made MATLAB code (adapted from a previously published code³⁵) to obtain the coordinates of the droplet contour (segmentation) and measure the in-plane curvature $\kappa(s)$ along the droplet contour, with s being the arc length along the contour. We first applied a Gaussian low-pass filter on the original raw image. The drop shape was identified from the filtered image using active contour segmentation to generate a mask. The edge was located by convolving the mask with the Sobel operator. The locations of edge pixels were then converted to ordered polar coordinates, which were smoothed using a moving average filter with a span of 5 pixels. We resampled the coordinates at even spacing using shape-preserving piecewise cubic interpolation. The curvature $\kappa(s)$ along the droplet contour was obtained by cubic fitting of edge coordinates over a small neighbourhood of each point along the contour.

Measurement of stresses associated with deviations from the ellipsoidal droplet deformations. To measure the value of stresses associated with deformation modes of higher order than the ellipsoidal deformation mode (droplet deformations at length scales smaller than the droplet size), we used a similar procedure to the one described²² to obtain cellular stresses from 2D confocal sections. Maximal and minimal values of the curvature along the droplet contour were spaced on average by the measured cell size (Fig. 1h, j), confirming that these deformations are associated with spatial variations in stresses occurring at the cell scale. In brief, we first segmented the 2D droplet shape and obtained the in-plane curvature $\kappa(s)$ along the droplet contour, as described above. We then calculated the deviations $\delta\kappa(s) \equiv \kappa(s) - \kappa_e(s)$ of the curvature along the contour from the curvature κ_e of the elliptical deformation mode. We fitted an ellipse to the contour coordinates using the MATLAB function `EllipseDirectFit` (written by N. Chernov based on a previous algorithm) to determine the elliptical curvature κ_e , and then calculated directly $\delta\kappa(s)$ along the droplet contour. We then detected the maxima and minima of $\delta\kappa$ along the contour and defined the amplitude of curvature deviations from the elliptical mode as the difference between a consecutive maximum and minimum of curvature, $\delta\kappa^{\max}$ and $\delta\kappa^{\min}$, respectively, along the droplet contour, namely $\Delta\kappa_C \equiv \delta\kappa^{\max} - \delta\kappa^{\min}$. We also defined the maximum amplitude of curvature deviations from the elliptical deformation mode, $\Delta\kappa_C^{\max} \equiv \delta\kappa^{\text{absmax}} - \delta\kappa^{\text{absmin}}$, where $\delta\kappa^{\text{absmax}}$ and $\delta\kappa^{\text{absmin}}$ correspond to the absolute maximum and absolute minimum in $\delta\kappa$ along the droplet contour. The average and maximal values of the stresses associated to cell-sized deviations from the ellipsoidal mode, $\sigma_C^{\text{average}}$ and σ_C^{\max} respectively, are given by $\sigma_C^{\text{average}} = \langle 2\gamma\Delta\kappa_C \rangle$ (average, indicated by the angled brackets, was calculated over multiple consecutive maxima and minima for a single droplet and also over multiple droplets in the same region of the tissue in different embryos) and $\sigma_C^{\max} = \langle 2\gamma\Delta\kappa_C^{\max} \rangle$ (average was done over multiple droplets in the same region of the tissue in different embryos). As previously noted²², this calculation assumes no major structural anisotropies in the tissue.

To measure the length scale associated with shape deviations from the elliptical deformation mode, we obtained the locations of maxima and minima of $\delta\kappa(s)$ along a droplet's contour, and calculated the contour distance λ between consecutive maxima and minima (Fig. 1h).

Measurement of persistence in supra-cellular (tissue-level) stresses. To quantify the persistence timescale of tissue-level stresses, we imaged equatorial sections of magnetic droplets (no applied magnetic fields, as described above) at time intervals of one second for up to 30 min. We fitted ellipses to droplet sections at each time point and reported the droplet aspect ratio as a function of time. We observed no substantial change in the average droplet aspect ratio over the course of 30 min, indicating that tissue-level stress anisotropies persist in the tissue for longer than at least 30 min.

Measurement of persistence in cell-scale stresses. To obtain the persistence of cell-scale stresses, we performed time-lapses of ferrofluid droplets (no magnetic actuation) inserted in the MPZ tissue at 2.5-s time intervals, for 10 to 16 min. We segmented those droplets at each time point and obtained the in-plane curvature $\kappa(s, t)$ along the contour (parameterized with the contour length s), as well as the curvature deviations from the elliptical droplet deformation, $\delta\kappa(s, t)$, as described above. For each time-lapse, we resampled $\delta\kappa(s, t)$ at equal angular (θ) spacing, to obtain $\delta\kappa(\theta, t)$. We then calculated the temporal autocorrelation $C_{\delta\kappa}(\tau)$, namely

$$C_{\delta\kappa}(\tau) = \frac{\langle (\delta\kappa(\theta, t + \tau) - \langle \delta\kappa(\theta, t + \tau) \rangle_{\theta, t}) (\delta\kappa(\theta, t) - \langle \delta\kappa(\theta, t) \rangle_{\theta, t}) \rangle_{\theta, t}}{\sqrt{\langle (\delta\kappa(\theta, t + \tau) - \langle \delta\kappa(\theta, t + \tau) \rangle_{\theta, t})^2 \rangle_{\theta, t}} \sqrt{\langle (\delta\kappa(\theta, t) - \langle \delta\kappa(\theta, t) \rangle_{\theta, t})^2 \rangle_{\theta, t}}}$$

We reported the value of the timescale $\tau_{1/2}$ for which $C_{\delta\kappa}(\tau)$ drops below 0.5 (or $C_{\delta\kappa}(\tau_{1/2}) = 0.5$), which corresponds to its half-life and provides a measure of the timescale of correlation loss or, equivalently, the persistence timescale of cell-scale droplet deformations.

Measurement of average cell size. Cell size was calculated using the polygonal drawing tool of ImageJ to outline the cells of each region and calculate the area of each selection. The average cell diameter was obtained by the diameter of a circle with the same area as that measured from the polygonal cell shape. The average cell size in a given region of the tissue was obtained from the ensemble average of 100 cells (approximately 35 in each region, A-PSM, P-PSM and MPZ).

Measurement of cell shape and nuclear anisotropy. Maximum intensity projections of 3 consecutive planes of a confocal stack, spanning in total $6\ \mu\text{m}$ in z , were subjected to thresholding and converted to binary images using ImageJ. Subsequently, the ImageJ plug-in 'Fill Holes' was applied and ellipses were fitted to the H2B-labelled nuclei. Nuclear anisotropy was given by the aspect ratio of the fitted ellipse. To quantify the orientation of nuclear elongation, we used ImageJ to measure the angle between the axis of nuclear elongation (given by the long axis of the fitted ellipse) and the anteroposterior axis. The direction of the anteroposterior axis was revealed by the notochord in the images. To obtain cell shape anisotropy, we first measured the length of cell–cell contacts oriented within 18° degrees of the anteroposterior and mediolateral axes. To do so, we used Tissue Analyzer (see 'Spatial and temporal autocorrelation of cell–cell contact length') to obtain the cell–cell contact lengths versus their orientation in 2D confocal sections of different regions.

Magnetic actuation of ferrofluid microdroplets. Actuation of ferrofluid droplets was done with the same equipment and methods as described²¹. In brief, we deformed a ferrofluid droplet previously injected in the tissue using an applied, uniform and constant magnetic field (as previously described²¹), thereby generating a controlled local strain in the tissue. We applied a uniform and constant magnetic field for 15 min, as this time is considerably longer than all measured stress relaxation timescales in the tissue²¹ (Supplementary Fig. 2b). Ferrofluid droplets deform into ellipsoids when a uniform, constant magnetic field is applied^{21,36}. We monitored the time evolution of the droplet deformation and measured the initial (b_0), maximal (b_M) and final (b_F) droplet semi-axis in the direction of the applied magnetic field (Fig. 2a, b). The final droplet semi-axis (b_F) was determined from the asymptotic deformation of the droplet after relaxation following the removal of the applied magnetic field. We defined the droplet strain as $\varepsilon \equiv (b - R)/R$. With this definition, the initial strain ε_0 , the maximum applied strain ε_M and the final strain ε_F are given by $\varepsilon_0 \equiv (b_0 - R)/R$, $\varepsilon_M \equiv (b_M - R)/R$ and $\varepsilon_F \equiv (b_F - R)/R$, respectively. We applied maximal mechanical strains in the 50–150% range, depending on the droplet's interfacial tension and the magnitude of the applied magnetic field.

Measurement of local yield stress with ferrofluid microdroplets. To measure yield stress, we applied strong magnetic fields to generate ellipsoidal droplet deformations larger than at least one cell size, as in these conditions the tissue flows irreversibly. Upon removal of the magnetic field, the capillary stress of the droplet pulls the droplet back towards the undeformed, spherical droplet shape (Fig. 2c). The capillary stress, σ_c , is the mechanical normal stress that tries to restore the spherical droplet shape because of the presence of an interfacial tension and depends on how deformed the droplet is: the larger the droplet deformation, the larger the capillary stress, as its value at a given point of the droplet's surface is proportional to the mean curvature H of the droplet at that point. Starting from the maximal droplet deformation induced by the applied magnetic field (maximal strain, ε_M), the capillary stresses driving droplet relaxation are large and become smaller as the droplet shape relaxes towards the sphere. In the absence of a yield stress in the tissue, the droplet would relax to the spherical shape. However, the presence of a yield stress halts the relaxation of the droplet at the aspect ratio for which the capillary stress is equal to the yield stress (Fig. 2c), preventing further droplet relaxation. At this point, the yield stress balances the droplet's capillary stresses (Fig. 2c). Since the interfacial tension of the droplet is measured in situ, the final aspect ratio of the arrested droplet serves as a direct readout of the yield stress. Importantly, we applied the magnetic field in the perpendicular direction to the droplet pre-deformation, so that the final aspect ratio in the direction of the magnetic field could not be attributed to the supracellular tissue anisotropy in mechanical stress that deformed the droplet initially. Moreover, to allow accurate comparisons between the measured anisotropic stresses and yield stress measurements, we defined the anisotropic yield stress as the anisotropic stress that results from the residual droplet deformation that remains after droplet relaxation is halted by the yield stress (Fig. 2c). The anisotropic yield stress is given by $\sigma_Y = 2\gamma(H_{b,r} - H_{a,r})$, where $H_{b,r}$ and $H_{a,r}$ are the mean curvatures of the residual droplet shape (deformation) at the intersection of the principal axis with the ellipsoid. Since the residual droplet shape is a prolate spheroid with long and short semi-axes b_r and a_r , respectively, $H_{b,r}$ and $H_{a,r}$ read $H_{b,r} = b_r/a_r^2$ and $H_{a,r} = 1/2a_r + a_r/(2b_r^2)$. The interfacial tension of each droplet was measured directly in situ, within the developing embryo, as previously described²¹. Using the measured value of

the interfacial tension for each droplet and the mean curvatures of the residual droplet deformation, we obtained the value of the yield stress, σ_Y .

The measured values of the yield stress did not depend on the extent of droplet deformation before starting the relaxation process (maximal droplet deformation or strain, ε_M ; Fig. 2a, b), ruling out potential memory effects that could occur in hysteretic plastic materials (Extended Data Fig. 10).

Measurement of volume fraction of extracellular space. To obtain the volume fraction of the extracellular space, we injected dextran–Alexa Fluor 488 (10,000 MW) in the MPZ of 9-somite stage embryos. After 30–45 min embryos were mounted and imaged. 3D reconstructions of different regions of the paraxial mesoderm were analysed in Imaris (Bitplane), using the surface reconstruction algorithm. The volumes encapsulated by the Alexa Fluor 488 segmented surfaces were then divided by the total volume analysed. Injection of dextran–Alexa Fluor 488 in the PSM of 9-somite stage embryos led to consistent results. The extracellular spaces are also visible by fluorescently labelling cell membranes (Extended Data Fig. 4), but the quantification of extracellular spaces was considerably more difficult and less accurate in this case.

Theoretical predictions of yield stress from volume fraction of extracellular space.

To obtain the predicted values of yield stress (gradient in wild type, Fig. 2f, and change between *cdh2* mutants and controls, Fig. 2j) from the measured values of the volume fraction ϕ of extracellular space, we assumed the tissue to behave as a 3D disordered monodispersed foam¹ for which the yield stress is given by $\sigma_Y = \sigma_Y^0(\phi_C - \phi)^2$, where ϕ_C is the critical volume fraction ($\phi_C = 0.36$ for 3D disordered monodispersed foams¹) and σ_Y^0 is a stress scale that depends on the interfacial tension of bubbles and their size. We assumed the magnitude of cell–cell contact tensions (playing the role of the bubble interfacial tension in an aqueous foam) to be constant along the anteroposterior axis, as our measurements indicated that cell-scale stresses are largely uniform in the tissue. The cell size was also assumed to be constant, because our measurements indicate no significant changes along the anteroposterior axis (Extended Data Fig. 9). The comparison between the experimental values of the yield stress and the predicted ones has no adjustable parameters (Fig. 2f, j).

Measurement and analysis of 3D cellular movements. To track cellular movements in 3D, both in control and *cdh2* mutant embryos, we first fluorescently labelled cellular nuclei by injecting H2B-mRFP mRNA into one-cell stage embryos. At the 10-somite stage, embryos were manually dechorionated and mounted in 0.8% low melting point agarose in a 35-mm glass-bottom dish (MatTek Corporation). Each region (A-PSM, P-PSM and MPZ) was imaged at 25°C using a $40\times$ water immersion objective (LD C-Apochromat 1.1W, Carl Zeiss Inc.), with z -stacks acquired at 1-min intervals for 30 min total. z -stacks were cropped to include only cells corresponding to the region of interest. Loss of signal intensity over time was corrected with the 'EMBL Bleach Correction' plugin for ImageJ using the histogram matching method. The 3D nuclei positions \mathbf{r}_i (with i specifying the nuclei) at each time point were determined using the Imaris (Bitplane) automated spot detection. We then used the Imaris (Bitplane) tracking algorithm to obtain the trajectories of the nuclei, $\mathbf{r}_i(t)$. For subsequent analysis, the results were filtered by track duration to include only those tracks that were maintained over the full course of the time lapse. We then used the nuclei 3D trajectories $\mathbf{r}_i(t)$ as a proxy for the trajectories of the cells. To eliminate the effect of translational and rotational motions of the whole tissue/embryo, we analysed the relative cell distances $\mathbf{r}_{ij} \equiv \mathbf{r}_i - \mathbf{r}_j$. We evaluated the mean-square displacement of the relative distances, or mean-square relative displacements (MSRD), from their initial values, that is, $\text{MSRD}(t) \equiv \langle (\mathbf{r}_{ij}(t) - \mathbf{r}_{ij}(0))^2 \rangle$, for (i, j) pairs, which were nearest neighbours at $t = 0$. We evaluated the normalized MSRD(t), namely $(\text{MSRD}(t) - \text{MSRD}(t = 0))/d^2$ (with d being the average cell size), for every (i, j) nearest pair in each region (A-PSM, P-PSM, and MPZ) of a given embryo, and averaged the results over all the pairs in the same region for each embryo.

Spatial and temporal autocorrelation of cell–cell contact length. To determine the characteristic timescale of changes in cell–cell contact lengths, we acquired a single confocal section of embryos injected with memGFP mRNA for 30 min with time resolution of 5 s. We detected the location of cell vertices and cell–cell contact lengths in the images using Tissue Analyzer³⁷ (formerly known as Packing Analyzer), a Fiji plugin capable of segmenting 2D tissue images and quantifying cell–cell contact lengths and position of vertices over time. For each embryo, we segmented a region of interest (ROI) in the PSM or tailbud (MPZ) for which cell–cell contacts were trackable over a 400 s time period. We then used the Tissue Analyzer package to obtain the contour length of cell–cell contacts and the (x, y) positions of the vertices. Using the timeseries of the contour length of cell–cell contacts and the (x, y) positions of the vertices, we calculated the temporal autocorrelation function, namely

$$C_\ell(\tau) \equiv \frac{\langle (\ell(t + \tau) - \langle \ell(t + \tau) \rangle)_i (\ell(t) - \langle \ell(t) \rangle)_i \rangle_t}{\sqrt{\langle (\ell(t + \tau) - \langle \ell(t + \tau) \rangle)_i^2 \rangle_t} \sqrt{\langle (\ell(t) - \langle \ell(t) \rangle)_i^2 \rangle_t}}$$

where ℓ is the cell–cell contact length. To reduce numerical errors, each average $\langle \bullet \rangle_t$ was evaluated with the data in a time interval $(0, T - \tau)$, with T being the duration of the experiment, and $\langle \bullet \rangle_{t+\tau}$ was evaluated with the data in the time interval (τ, T) . We analysed each ROI and obtained the correlation function for each embryo separately. The obtained autocorrelation functions were nearly exponential in all cases (Fig. 3c). We therefore obtained the characteristic timescale of the decay of the correlation function by fitting an exponential function. The reported characteristic timescale was obtained from a weighted average of the timescales measured in different embryos, with the weights being the inverse of the variance. To obtain the spatial cross-correlation, a similar analysis was done, but correlating instead every two pairs of cell–cell contact lengths in the region of the analysed tissue. The distance between cell–cell contact pairs was given by the norm of the vector connecting the midpoints of the two cell–cell contacts (Fig. 3d).

Normalized frequency (distribution) of cell–cell contact length fluctuations and energy landscape of neighbour exchanges. We obtained the distribution of normalized cell–cell contact lengths, and the corresponding dimensionless energy landscape, which characterizes neighbour exchanges. For the ROIs described above (in ‘Spatial and temporal autocorrelation of cell–cell contact length’), we selected all cell–cell contacts that persist for at least 100 s, and obtained their time-dependent normalized lengths $\ell/\bar{\ell}$, where ℓ is the cell–cell contact length at a given time point and $\bar{\ell}$ is the average length of each individual cell–cell contact over time. We then combined all the values of $\ell/\bar{\ell}$ for all times and cell–cell contacts into a single, normalized frequency distribution $p(\ell/\bar{\ell})$, which peaks around 1 (Fig. 3e). Assuming that the cell–cell contact length fluctuations are probing an energy landscape that varies much more slowly (equilibrium approximation) before they undergo neighbour exchanges, we obtained the dimensionless energy landscape of neighbour exchanges by calculating $-\ln[p(\ell/\bar{\ell})]$ (Extended Data Fig. 7). Since the statistics of cell–cell contact length fluctuations is small for values close to 0 (that is, close to a neighbour exchange event), this method provides the dimensionless energy landscape close to the bottom of the energy landscape but does not allow to obtain the size of the energy barrier.

Measurement of neighbour exchange rates and orientation. To quantify the neighbour exchange rate in the tissue, we analysed 2D confocal sections of untreated embryos and blebbistatin-treated embryos, both in the PSM and MPZ. We counted the number of cell–cell contacts that underwent a neighbour exchange in the plane of observation. We obtained the neighbour exchange rate per cell by normalizing the number of measured neighbour exchanges in the observed region with the average cell area and observation time (30 min). To obtain their orientation, we measured the angle between the cell–cell contact and the direction of the anteroposterior axis before neighbour exchange. The direction of the anteroposterior axis was revealed by the notochord in the images.

Inhibition of myosin activity. Myosin activity was inhibited by adding in the embryo medium blebbistatin (Tocris) at the final concentration of 100 μM . Treatments were started at the four-somite stage.

Measurement of cell density. To measure cell number density (number of cells per unit volume), we performed confocal 3D scans of the different regions (MPZ, P-PSM and A-PSM) in embryos with fluorescently labelled nuclei (same procedure as described above in ‘Measurement and analysis of 3D cellular movements’). We then used Imaris software to detect the location of the nuclei in the tissue (approximately 300 cells per imaged volume). After obtaining the 3D coordinates of each nucleus in a volume, we defined a box of smaller size than the full scanned volume and measured the number of nuclei in the box. Moving the small box inside the

originally scanned volume, we obtained statistics for the density measurement. This procedure was repeated with varying box sizes to ensure that the measured value of the density did not depend on the box size.

Actin staining. Ten-somite stage embryos were dechorionated and fixed in 4% paraformaldehyde for 1.5 h at room temperature. Embryos were then washed 2–3 times with PBS. One drop of ActinRed was added to the PBS solution (500 μl) containing the embryos. After 30 min, the embryos were washed with new PBS 3 times and prepared for imaging as described above.

Simulations. We simulated the physical expansion of a 2D dorsal–ventral projection of the PSM and MPZ tissues at supracellular length scales and developmental time scales using finite element methods, accounting for the fluid-to-solid jamming transition (or its absence, depending on the simulation) and the entrance of cells to the MPZ from dorsal medial tissues (see Supplementary Note 1 for details). The fluid-to-solid jamming transition was simulated by a sharp increase in the tissue viscosity along the anteroposterior axis, with a diverging (very large) viscosity simulating the solid-like tissue state. The dynamics of the tissue is governed by fundamental physical laws, namely momentum conservation (which reduces here to local force balance because inertial terms are negligible for embryonic tissues) and mass balance. We solved the equations using COMSOL Multiphysics 5.3 software, starting from a semi-circular tissue shape and with a fixed rigid boundary at the most anterior end. The solutions provided the time evolution of the tissue shape and velocity field.

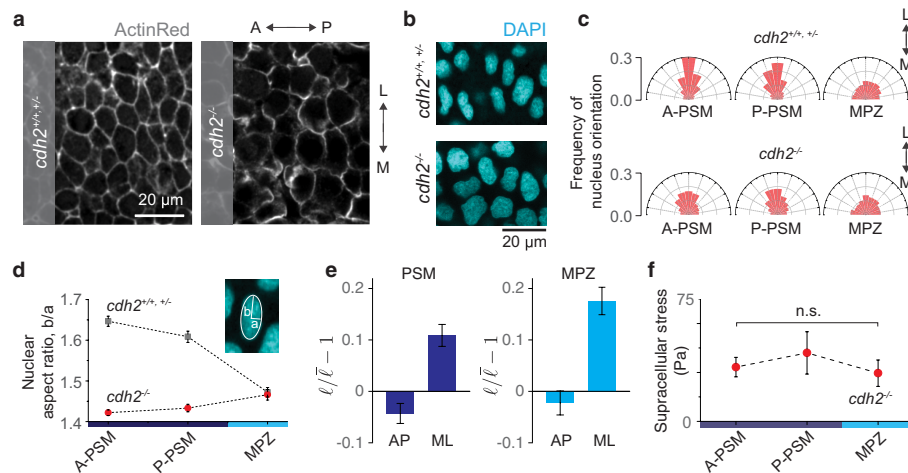
Statistics. In experiments involving zebrafish embryos, the sample size was chosen so that new data points would not significantly change the standard deviation. No samples were excluded from the analysis and the analysis of all the data was done by automated software to ensure full blinding and avoid biases in the analysis. No randomization of the data was used.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. A previously published version of the custom-made MATLAB code to analyse droplet deformations³⁵ is available on GitHub at <https://github.com/elijahshelton/drop-recon>. A modified version of that code to analyse 2D confocal sections of the droplet is available upon request.

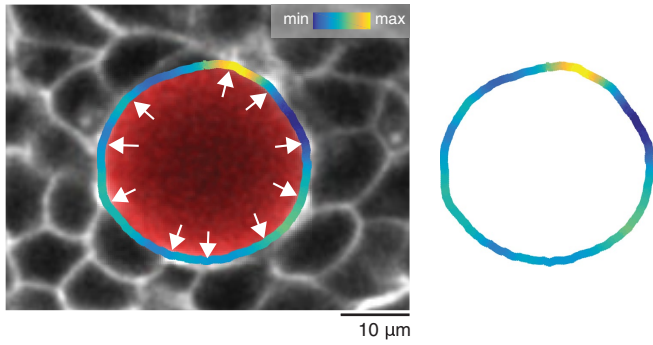
Data availability. Source Data for Figs. 1–3 and Extended Data Figs. 1, 5–10 are available in the online version of the paper.

30. Nüsslein-Volhard, C. & Dahm, R. *Zebrafish* (Oxford Univ. Press, Oxford, 2002).
31. Mosaliganti, K. R., Noche, R. R., Xiong, F., Swinburne, I. A. & Megason, S. G. ACME: automated cell morphology extractor for comprehensive reconstruction of cell membranes. *PLOS Comput. Biol.* **8**, e1002780 (2012).
32. Holtze, C. et al. Biocompatible surfactants for water-in-fluorocarbon emulsions. *Lab Chip* **8**, 1632–1639 (2008).
33. Sletten, E. M. & Swager, T. M. Fluorofluorophores: fluorescent fluorocarbon chemical tools spanning the visible spectrum. *J. Am. Chem. Soc.* **136**, 13574–13577 (2014).
34. Rallison, J. The deformation of small viscous drops and bubbles in shear flows. *Annu. Rev. Fluid Mech.* **16**, 45–66 (1984).
35. Shelton, E., Serwane, F. & Campas, O. Geometrical characterization of fluorescently labelled surfaces from noisy 3D microscopy data. *J. Microsc.* **269**, 259–268 (2017).
36. Rowghanian, P., Meinhart, C. D. & Campas, O. Dynamics of ferrofluid drop deformations under spatially uniform magnetic fields. *J. Fluid Mech.* **802**, 245–262 (2016).
37. Aigouy, B. et al. Cell flow reorients the axis of planar polarity in the wing epithelium of *Drosophila*. *Cell* **142**, 773–786 (2010).

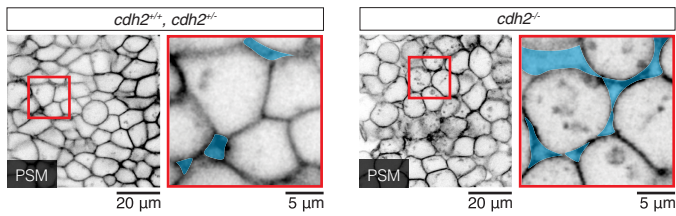


Extended Data Fig. 1 | Loss of anteroposterior gradients of supracellular stresses and cell and nuclear shape anisotropy in N-cadherin mutants. **a**, ActinRed staining of F-actin in the PSM of control (*cdh2*^{+/+} and *cdh2*^{+/-}) and mutant (*cdh2*^{-/-}) embryos at the 10-somite stage. Cell shapes are visibly elongated along the mediolateral (ML) direction in control embryos. Cell shape anisotropy is largely lost in *cdh2*^{-/-} embryos. **b**, DAPI staining showing higher nuclear mediolateral elongation in the PSM of control embryos compared to *cdh2* mutants. **c**, Frequency of nuclear major axis orientations in the MPZ and PSM (A-PSM and P-PSM). In control (*cdh2*^{+/+} and *cdh2*^{+/-}) embryos, nuclei in the PSM are elongated along the mediolateral direction, whereas nuclei are oriented randomly in the MPZ. The observed nuclear anisotropy along the mediolateral direction in the PSM of control embryos is decreased in *cdh2* mutants. **d**, A posterior-to-anterior increase in the extent of nuclear elongation (quantified by the nuclear aspect ratio; see inset and Methods) is observed in control (*cdh2*^{+/+} and *cdh2*^{+/-}) embryos. No anteroposterior gradient in the extent of nuclear elongation (aspect ratio) is observed in *cdh2* mutants (*cdh2*^{-/-}). For **c** and **d**: control embryos, *n* = 695 (A-PSM), 752 (P-PSM), 732 (MPZ) nuclei, obtained in 6 embryos per region;

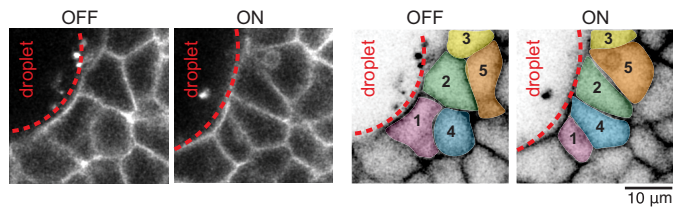
cdh2^{-/-}: *n* = 833 nuclei from 5 embryos (A-PSM), *n* = 538 nuclei from 6 embryos (P-PSM), *n* = 336 nuclei from 4 embryos (MPZ). Mean ± s.e.m. **e**, Relative change of cell-cell contact length along the anteroposterior axis and the mediolateral axis (*n* = 6,427 and 4,319 cell-cell contacts for PSM and MPZ, respectively, from 5 embryos). Cell junctions are longer along the mediolateral axis compared to the anteroposterior axis, both in the PSM and MPZ. Mean ± s.e.m. **f**, Supracellular stresses are uniform along the anteroposterior axis in *cdh2* mutant embryos (*n* = 5 (A-PSM), 5 (P-PSM), 12 (MPZ)). Mean ± s.e.m.; Mann-Whitney *U*-test. The observed posterior-to-anterior increase in both supracellular stresses and nuclear elongation in control embryos (**d** and Fig. 1e), and the loss of both such gradients in *cdh2* mutants (**d** and **f**), indicate the existence of a N-cadherin-dependent, posterior-to-anterior increase in supracellular stresses, consistent with a posterior-to-anterior increase in mediolateral constriction. Importantly, if the observed thinning of the body axis was caused by pulling forces from the MPZ on the PSM, as previously proposed¹⁸, both cells and nuclei would be elongated along the anteroposterior axis.



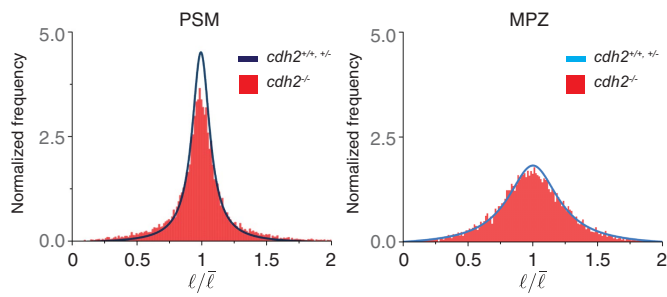
Extended Data Fig. 2 | Curvature changes along the droplet contour correlate with the locations of cell–cell contacts surrounding the droplet. Confocal section of a ferrofluid droplet (red) in the MPZ of a *Tg(actb2:MA-citrine)* embryo. The measured curvature values along the detected droplet contour are shown (colourcoded as in Fig. 1h) overlaid with the confocal image (left) and without it (right). White arrows point to locations of cell–cell contacts of cells surrounding the droplet, which correlate with maxima and minima of droplet curvature, consistent with the distance between maxima and minima being approximately the cell size (Fig. 1j).



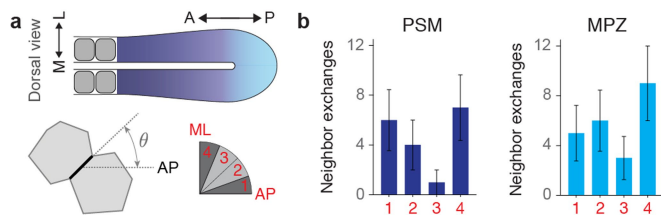
Extended Data Fig. 3 | Increase of extracellular spaces and cell rounding in *cdh2* mutants. 2D confocal sections (inverted) of control (*cdh2*^{+/+} and *cdh2*^{+/-}) and *cdh2*^{-/-} Tg(*actb2*:MA-citrine) embryos showing an increase in extracellular space (cyan), as well as more cell rounding, in the PSM tissue of the mutant embryos.



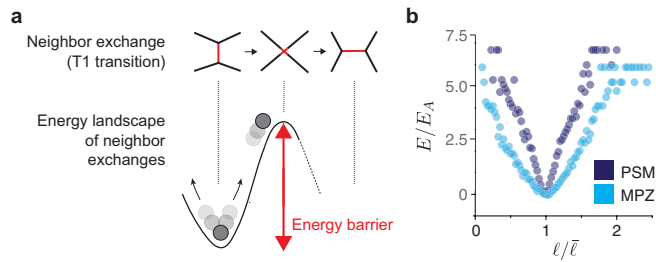
Extended Data Fig. 4 | Example of neighbour exchanges induced in the tissue upon droplet actuation with a magnetic field. Confocal section showing the spatial arrangements of cells in the neighbourhood of a magnetically responsive droplet both in the absence of magnetic field (OFF) and after applying a magnetic field (ON) for 15 min (left). Several cell rearrangements are observed to be induced by droplet actuation (right). Some of the cells undergoing neighbour exchanges are coloured and numbered to highlight the rearrangements. Tg(*actb2*:MA-citrine) embryos were used to visualize cell membranes.



Extended Data Fig. 5 | Distribution of cell–cell contact length fluctuations in *cdh2* mutants. Normalized frequency (distribution) of cell–cell contact length fluctuations in the PSM and MPZ of *cdh2* mutants (red bars) compared to the control (violet and light blue lines). For PSM and MPZ, $n = 13,212$ and $13,634$ cell–cell contacts obtained from 5 and 4 embryos, respectively.



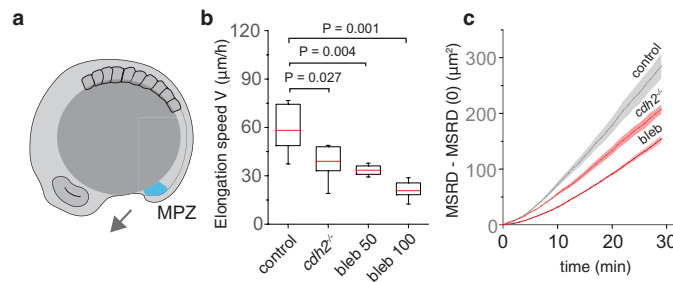
Extended Data Fig. 6 | Orientation of neighbour exchanges in the MPZ and PSM. **a**, Sketch of a dorsal view of the elongating body axis, with the anteroposterior and mediolateral directions defined (top). Sketch showing the orientation of a cell-cell contact (thick black line) before undergoing a neighbour exchange (bottom left). The angle θ corresponds to the angle between the cell-cell contact before undergoing the neighbour exchange and the anteroposterior axis (bottom). Four equal bins are defined (bin 1: $0 < \theta < 22.5^\circ$; bin 2: $22.5^\circ < \theta < 45^\circ$; bin 3: $45^\circ < \theta < 67.5^\circ$; bin 4: $67.5^\circ < \theta < 90^\circ$) between the anteroposterior and mediolateral orthogonal directions (bottom right). **b**, Frequency of neighbour exchanges along different angular regions ($n = 18$ in 4 embryos for PSM and $n = 23$ in 3 embryos for MPZ, with n being the number of neighbour exchanges analysed). Data are mean \pm s.d. Neighbour exchanges are largely randomly oriented in the MPZ. In the PSM, neighbour exchanges occur predominantly along either the mediolateral direction or along the anteroposterior axis, with neighbour exchanges occurring slightly less frequently for angles in between these orthogonal orientations. The more frequent occurrence of neighbour exchanges along the anteroposterior and mediolateral axes in the PSM is consistent with the measured directions and extent of ellipsoidal droplet deformation (Fig. 1f), as the persistent and larger supracellular stresses in the PSM may bias neighbour exchanges in these directions. Since neighbour exchanges occur equally frequently along the mediolateral and anteroposterior directions in the PSM, and are uniformly oriented in the MPZ, our results indicate no systematic alignment of neighbour exchanges along a single spatial direction that could potentially contribute to the elongation of the body axis.



Extended Data Fig. 7 | Energy landscape of neighbour exchanges.

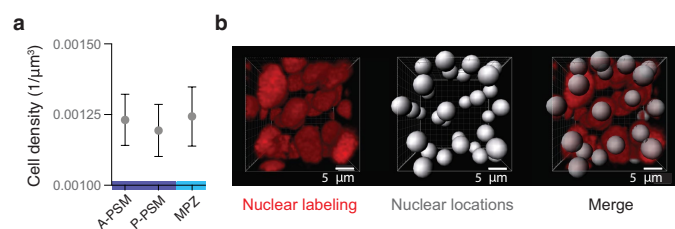
a, Schematic of key cellular configurations throughout a neighbour exchange and associated energy landscape. Changing neighbours requires overcoming an energy barrier. Large enough, active cell–cell contact length fluctuations enable neighbour exchanges. **b**, Measured energy

landscape, E , for PSM and MPZ regions, normalized with the energy scale E_A associated with cell–cell contact activity or effective temperature energy scale, namely $E_A = k_B T_{\text{eff}}$, where k_B is the Boltzmann constant and T_{eff} is the effective temperature. $n = 6,969, 7,896$ cell–cell contacts obtained from 3, 4 embryos for PSM, MPZ, respectively.

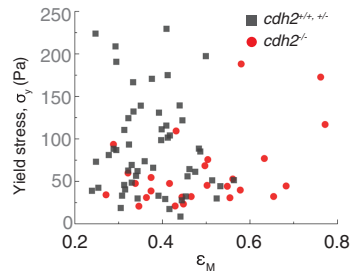


Extended Data Fig. 8 | Dependence of posterior axis elongation speed and relative cellular movements in the MPZ on N-cadherin and non-muscle myosin-II activity. **a**, Sketch of a 10-somite stage embryo highlighting the mesodermal progenitor zone (MPZ, cyan) and the direction of posterior elongation (arrow). **b**, Comparison of posterior body elongation speeds between control ($n = 6$), $cdh2$ mutants ($n = 7$), and blebbistatin-treated embryos ($n = 6$ for 50 μM and $n = 7$ for 100 μM).

Box plots representing median (red line) and second and third quartiles. Error bars indicate 95% confidence interval. Mann–Whitney U -test. **c**, Mean square relative displacement of cells in the MPZ region of control ($n = 2,523$ analysed cell pairs from 6 embryos), $cdh2^{-/-}$ ($n = 1,154$ analysed cell pairs from 4 embryos) and blebbistatin-treated embryos ($n = 2,026$ analysed cell pairs from 4 embryos).



Extended Data Fig. 9 | Cell density is uniform along the anteroposterior axis. **a**, Measured cell number density (cells per unit volume) in the MPZ, P-PSM and A-PSM. Mean \pm s.e.m. Cell density does not vary significantly along the anteroposterior axis (within the 10% accuracy of our 3D measurements; Methods). **b**, 3D reconstructions of confocal stacks showing nuclei labelled with H2B::RFP, detected nuclei positions, and composition of both. Cell density was measured using 3D data of nuclear positions in the different regions ($n = 7,866, 7,214, 11,537$ detected cells in 694, 694, 833 defined boxes in 5, 5, 6 embryos, in A-PSM, P-PSM, MPZ, respectively; Methods).



Extended Data Fig. 10 | Yield stress values do not depend on the extent of droplet deformation before droplet relaxation. Measured values of the yield stress plotted against the maximal droplet deformation (maximal applied strain, ϵ_M ; Fig. 2a, b) before starting droplet relaxation. The measured yield stress values do not correlate with the maximal applied strain, neither in control ($cdh2^{+/+}$ and $cdh2^{+/-}$, grey dots, $n = 53$ embryos) or mutant embryos ($cdh2^{-/-}$, red dots, $n = 27$ embryos). Correlation coefficient, $r = -0.34$ (control), $r = -0.04$ ($cdh2^{-/-}$).

Tracing HIV-1 strains that imprint broadly neutralizing antibody responses

Roger D. Kouyos^{1,2,15,16*}, Peter Rusert^{1,15}, Claus Kadelka^{1,2,15}, Michael Huber¹, Alex Marzel^{1,2}, Hanna Ebner¹, Merle Schanz¹, Thomas Liechti^{1,13}, Nikolas Friedrich¹, Dominique L. Braun^{1,2}, Alexandra U. Scherrer^{1,2}, Jacqueline Weber¹, Therese Uhr¹, Nicolas S. Baumann¹, Christine Leemann^{1,2}, Herbert Kuster^{1,2}, Jean-Philippe Chave³, Matthias Cavassini⁴, Enos Bernasconi⁵, Matthias Hoffmann⁶, Alexandra Calmy⁷, Manuel Battegay⁸, Andri Rauch⁹, Sabine Yerly¹⁰, Vincent Aubert¹¹, Thomas Klimkait¹², Jürg Böni¹, Karin J. Metzner^{1,2}, Huldrych F. Günthard^{1,2,16*}, Alexandra Trkola^{1,16*} & The Swiss HIV Cohort Study¹⁴

Understanding the determinants of broadly neutralizing antibody (bNAb) evolution is crucial for the development of bNAb-based HIV vaccines¹. Despite emerging information on cofactors that promote bNAb evolution in natural HIV-1 infections, in which the induction of bNAbs is genuinely rare², information on the impact of the infecting virus strain on determining the breadth and specificity of the antibody responses to HIV-1 is lacking. Here we analyse the influence of viral antigens in shaping antibody responses in humans. We call the ability of a virus strain to induce similar antibody responses across different hosts its antibody-imprinting capacity, which from an evolutionary biology perspective corresponds to the viral heritability of the antibody responses. Analysis of 53 measured parameters of HIV-1-binding and neutralizing antibody responses in a cohort of 303 HIV-1 transmission pairs (individuals who harboured highly related HIV-1 strains and were putative direct transmission partners or members of an HIV-1 transmission chain) revealed that the effect of the infecting virus on the outcome of the bNAb response is moderate in magnitude but highly significant. We introduce the concept of bNAb-imprinting viruses and provide evidence for the existence of such viruses in a systematic screening of our cohort. The bNAb-imprinting capacity can be substantial, as indicated by a transmission pair with highly similar HIV-1 antibody responses and strong bNAb activity. Identification of viruses that have bNAb-imprinting capacities and their characterization may thus provide the potential to develop lead immunogens.

The capacity to evoke highly similar bNAb responses across vaccinees is crucial for an effective HIV-1 immunogen. Closely related HIV-1 strains may induce similar neutralization responses, as observations from mother-to-child transmission suggest³. To formally evaluate the virus-dictated heritability of antibody responses, we investigated the imprinting capacity of HIV-1 antibody responses within transmission pairs. We designed our study to address two central problems (Extended Data Fig. 1a, b). First, we investigated whether the same virus, when transmitted to two different people, induces similar binding and neutralizing antibody responses (imprints a similar antibody response). Second, we investigated how promising HIV-1 strains with superior bNAb-imprinting capacity can be identified.

On the basis of the Swiss 4.5K Screen^{4,5}, we established a large, adult transmission-pair cohort ($n = 303$ putative transmission pairs) with comprehensive information on HIV-1-binding and neutralizing antibody responses (Extended Data Fig. 1a). Extensive data on HIV-binding antibody reactivity encompassing IgG1, IgG2 and IgG3

reactivity with 13 antigens was available for all 606 patients from previous analyses⁵ (Supplementary Data 1). Neutralization activity was assessed against a 14 multi-clade virus panel (Extended Data Fig. 2a and Supplementary Data 1, 2) and evaluated by breadth and a cumulative neutralization score, reflecting potency and breadth across the analysed virus panel (Extended Data Fig. 2a–d). Overall, the neutralization activity in the transmission pair cohort showed the typical pattern seen in chronic infection⁴: the majority of patients displayed no or low neutralization activity (73% of patients had below 10% breadth).

We hypothesized that if virus-associated factors are important in determining antibody responses, HIV antibody response patterns should be similar in transmission pairs. Using the established 53 HIV-1 antibody parameters (14 neutralization and 39 binding antibody parameters), we conducted a systematic assessment of the HIV-1 antibody imprinting capacity in transmission pairs (Extended Data Fig. 2e).

We detected a significant, positive association of the transmitter and recipient neutralization responses to 7 of the 14 panel viruses (Fig. 1a and Extended Data Fig. 3a). The overall similarity of the neutralization fingerprint within pairs across the 14 panel viruses was assessed as average Spearman correlation ($\rho_{\text{Spearman-average}}$) of their neutralization activity (Fig. 1b). To determine the statistical significance of the observed similarity, we used shuffling approaches that randomly reassign recipients to transmitters, thus generating a distribution for the null-expectation of no association. Neutralization fingerprints in observed transmission pairs proved on average positively and significantly associated ($\rho_{\text{Spearman-average}} = 0.11$, $P_{\text{shuffling}} < 0.001$; Fig. 1b). To confirm the influence of the infecting virus, we estimated the heritability of antibody responses by two alternative methods adjusting for the influence of various host, viral and disease factors that are known to influence antibody responses^{4,5}. First, we restricted the shuffling to pairs with the same infection length, subtype and ethnicity (Fig. 1b). Second, we considered mixed-effect Tobit models adjusted for key drivers of HIV-1 antibody development (infection length, ethnicity, virus load and viral diversity) and bNAb specificity (HIV-1 *pol* subtype) (Fig. 1c). Both approaches confirmed a significant, within-pair correlation of neutralization (Fig. 1b, c). Although other, not yet defined, non-virus-associated factors common to both transmission partners may exist, our data strongly suggest that the infecting virus strain affects the development of neutralization responses. The effects (Fig. 1a, b) remained robust when restricting the analysis to pairs infected with subtype B virus, indicating that the effect is not driven by specific subtypes (Extended Data Fig. 4a, b).

¹Institute of Medical Virology, University of Zurich, Zurich, Switzerland. ²Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland. ³Clinique de La Source, Lausanne, Switzerland. ⁴Division of Infectious Diseases, University Hospital Lausanne, University of Lausanne, Lausanne, Switzerland. ⁵Division of Infectious Diseases, Regional Hospital Lugano, Lugano, Switzerland. ⁶Division of Infectious Diseases, Cantonal Hospital St. Gallen, St. Gallen, Switzerland. ⁷Division of Infectious Diseases, University Hospital Geneva, University of Geneva, Geneva, Switzerland. ⁸Division of Infectious Diseases, University Hospital Basel, University of Basel, Basel, Switzerland. ⁹Department of Infectious Diseases, University Hospital Bern, University of Bern, Bern, Switzerland. ¹⁰Laboratory of Virology, Division of Infectious Diseases, University Hospital Geneva, University of Geneva, Geneva, Switzerland. ¹¹Division of Immunology and Allergy, University Hospital Lausanne, University of Lausanne, Lausanne, Switzerland. ¹²Division of Infection Diagnostics, Department of Biomedicine-Petersplatz, University of Basel, Basel, Switzerland. ¹³Present address: Immunotechnology Section, Vaccine Research Center, NIAID, National Institutes of Health, Bethesda, MD, USA. ¹⁴A list of participants and their affiliations appears at the end of the paper. ¹⁵These authors contributed equally: Roger D. Kouyos, Peter Rusert, Claus Kadelka. ¹⁶These authors jointly supervised this work: Roger D. Kouyos, Huldrych F. Günthard, Alexandra Trkola. *email: roger.kouyos@usz.ch; huldrych.guenthard@usz.ch; trkola.alexandra@virology.uzh.ch

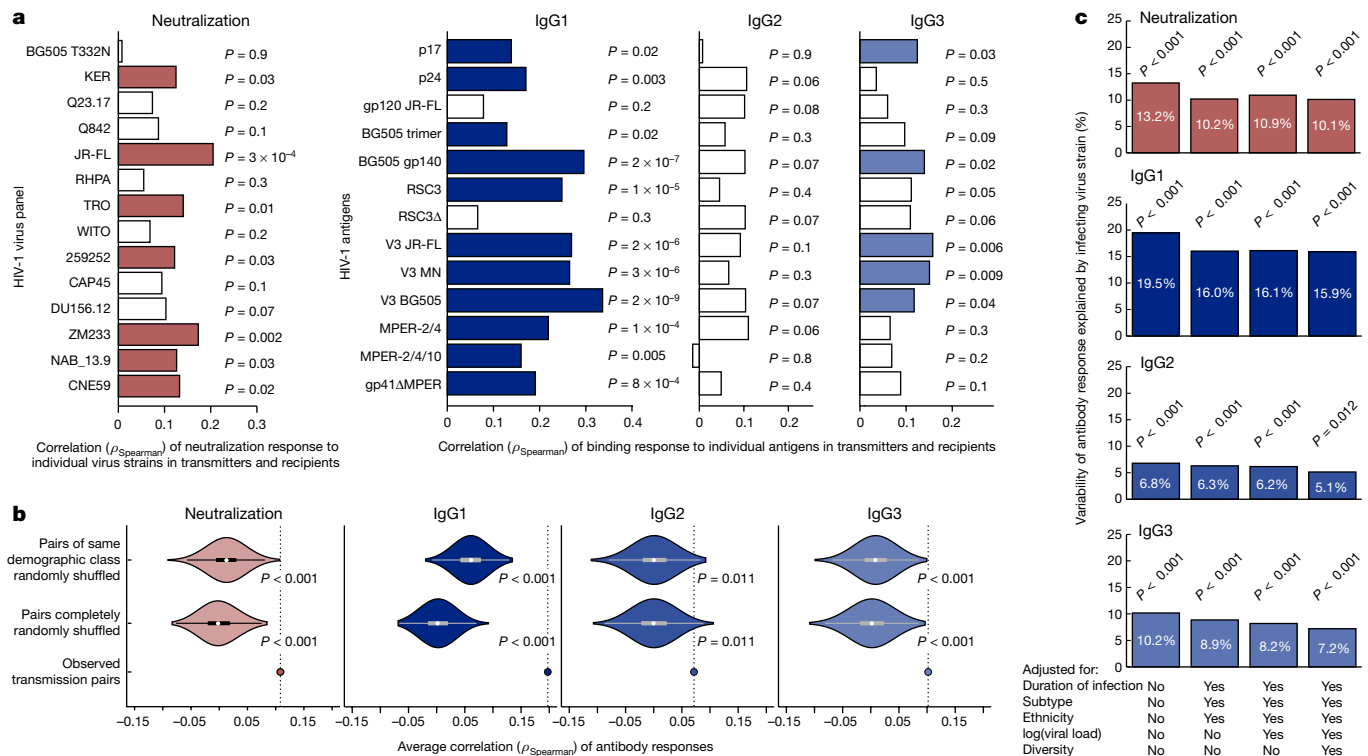


Fig. 1 | Similarity of neutralization and antibody-binding responses in transmission pairs. **a**, Spearman correlation of antibody responses in observed transmission pairs ($n = 303$; see also Extended Data Figs. 2e, 3a). Significant correlations (two-sided $P_{\text{Spearman}} < 0.05$) are coloured. **b**, Average Spearman correlation of antibody responses in observed transmission pairs ($n = 303$) compared to two alternative scenarios: (1) completely random reassignment of recipients to transmitters and (2) random reassignment of recipients to transmitters with the same demographics (subtype, ethnicity, and untreated infection length). Violins

(smoothed using a normal kernel) and one-sided P values were derived from 1,000 random reassignments of recipients to transmitters. Medians (white dots) and boxes spanning the interquartile range (IQR) range are shown. Each whisker extends to the most extreme value no more than $1.5 \times \text{IQR}$ from the box. **c**, Proportion of variability in antibody responses explained by the infecting virus, determined using unadjusted and differently adjusted mixed-effect Tobit regression models ($n = 303$). One-sided P values were derived from comparison with 1,000 random reassignments of recipients to transmitters.

Unravelling the quantitative contribution of the infecting virus to the neutralization response is of particular importance for vaccine development. The mixed-effect Tobit models revealed that, on average, 13.2% of the variability of the neutralization response can be explained by the infecting virus (Fig. 1c). Notably, various alternative models adjusted for cofactors of HIV-1 antibody induction—such as viral load or duration of infection—yielded similar results (9.3–13.8% neutralization variability explained by the virus; Extended Data Fig. 3b). Comparable results were also obtained when we restricted the analysis to the 184 pairs in which both individuals were infected for three or more years (neutralization heritability in Tobit models: 13.2% unadjusted, 9.3% fully adjusted) or when we used multiple imputation instead of a complete case analysis in the fully adjusted model (10.9% neutralization variability; Extended Data Fig. 3c).

HIV-1 antigen-binding activities were also significantly correlated within transmission pairs, but the degree of correlation differed considerably across antigens and IgG classes (Fig. 1a). IgG1 reactivity, the most prominent IgG response in HIV-1 infection⁵, displayed the highest similarity in transmission pairs (Fig. 1a, b), followed by IgG3 and IgG2 responses. Of note, IgG2 responses, which are only present at low levels during HIV-1 infection, showed no statistically significant similarity at the level of individual antigens. The average within-pair similarity across antigens was nevertheless significantly positive for all three IgG subclasses, with and without taking potentially confounding parameters into account (Fig. 1b). As observed for neutralization responses, the findings remained significant when the analysis was restricted to subtype B infection (Extended Data Fig. 4c, d). The effect of the infecting virus was comparable in magnitude to the effect on neutralization (19%, 7% and 10% for IgG1, IgG2 and IgG3 responses, respectively), with similar values observed after adjustment for analysed cofactors

of HIV-1 antibody induction (Fig. 1c). The identified influence of the infecting virus on the heritability of HIV-1 binding and neutralizing antibody responses was robust, as confirmed by additional sensitivity analyses (Extended Data Figs. 3b, 5a–c).

The overall effect sizes of the influence of virus genetics on neutralizing and binding antibodies that we report here must be considered as lower bound estimates, as neutralization breadth is generally low in HIV-1 infection. Furthermore, as our cohort setup did not allow assessing matched time points, transmission partners had experienced different lengths of virus replication and antibody response development at sample collection. Although the effects that we identified across the cohort were moderate, individual cases may substantially exceed the observed average antibody similarity. In support of this, two HIV-1 subtype B-infected elite neutralizers (top 1% of neutralizers identified in the Swiss 4.5K Screen⁴) formed one transmission pair (0.0058 HIV-1 *pol* genetic distance) that stood out in both potency and breadth (Fig. 2a, b). This pair was subsequently verified by their shared treating physician as a mixed ethnicity heterosexual couple in a stable partnership (T282 female (North African) to R282 male (Asian) transmission). Identification of transmission pairs and genetic subtype information in our cohort is based on *pol* sequence data collected early after diagnosis. We performed full-genome next-generation sequencing of plasma viruses from T282 and R282 at the time point analysed for antibody reactivity to rule out later superinfection. The obtained consensus full-genome sequences confirmed closely related subtype B viruses in T282 and R282 (Extended Data Fig. 6a). Because the analysed samples of T282 and R282 were collected 1.9 and 3.4 years after the putative transmission event, respectively, considerable genetic differences in Env between the pair were expected⁶. Env similarity in the pair was nevertheless evident, as highlighted by phylogenetic comparison,

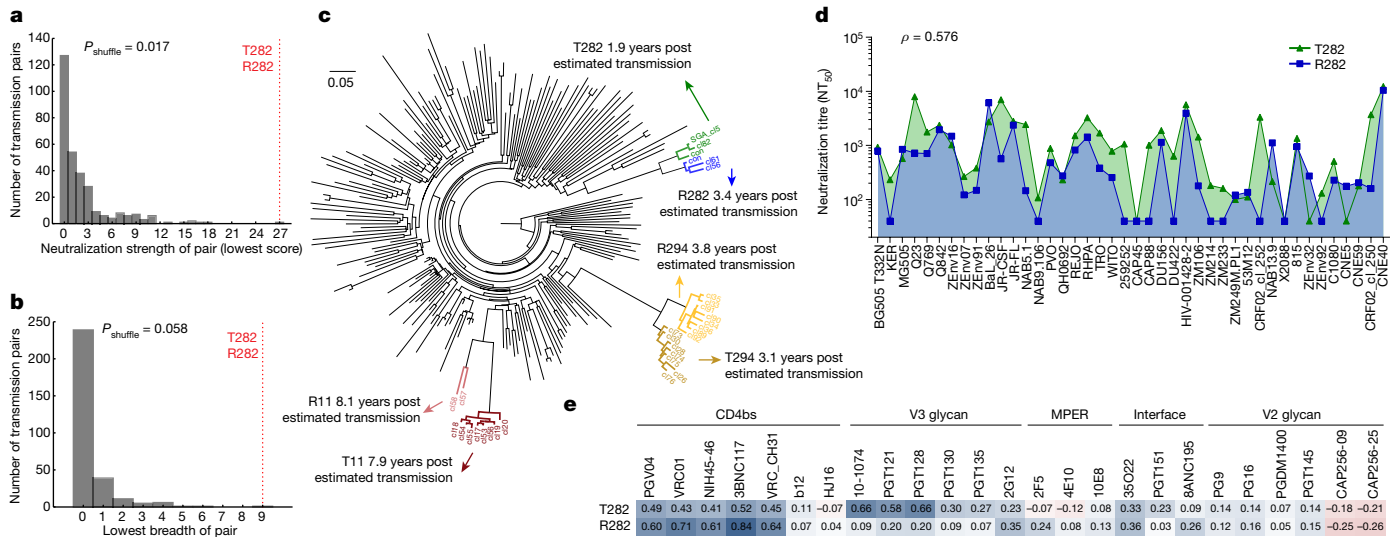


Fig. 2 | Identification of an elite-neutralizing transmission pair, T282 and R282. **a, b**, Lowest neutralization strength (**a**) and lowest breadth (**b**) of recipients and transmitters in observed transmission pairs ($n = 303$; red line: elite-neutralizing pair T282 and R282). One-sided P values were derived from comparison of maximum value with 10^6 random reassignments of recipients to transmitters. **c**, Phylogenetic analysis of Env sequences from pairs T282 and R282 (green and blue), T11 and R11

(red), and T294 and R294 (yellow), combined with 198 closely related Env background sequences (see Supplementary Methods and Extended Data Fig. 6). Years post estimated transmission at sampling indicated for transmission pairs. **d**, The 50% neutralization titres of plasma from T282 and R282 against the multi-clade virus panel ($n = 42$; $\rho_{\text{spearman}} = 0.576$). **e**, Neutralization fingerprint similarity (ρ_{spearman}) of plasma from T282 and R282 with 25 known bNABs.

validating the close relatedness of the viruses infecting the transmission partners (Fig. 2c). The same pattern was observed for two other transmission pairs (T11–R11 and T294–R294) analysed for comparison.

In line with reactivity to an originally highly related Env antigen, neutralization fingerprint analysis of plasma samples from T282 and R282 in a 42-virus panel confirmed highly similar neutralization responses in this pair ($\rho_{\text{spearman}} = 0.576$; Fig. 2d). Delineation of plasma bNAb

specificity by fingerprint analyses (Fig. 2e and Supplementary Data 2) and mutant-virus neutralization mapping (Extended Data Fig. 7a) showed that both partners developed a CD4 binding-site-directed bNAb plasma response and that additional V3 glycan bNAb activity was present in the transmitter.

The occurrence of a transmission pair with such pronounced neutralization strength and similarity in neutralization responses is

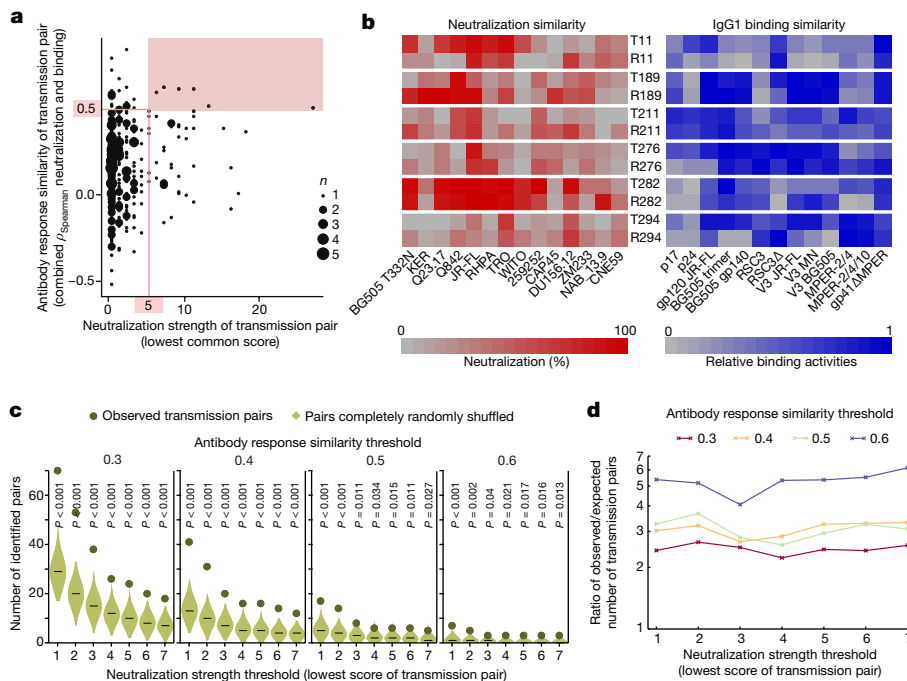


Fig. 3 | Systematic screen for virus strains with bNAb-imprinting capacity. **a**, Antibody response similarity versus neutralization strength of 303 transmission pairs (see Extended Data Fig. 8). **b**, Comparison of neutralization and IgG1-binding profiles for pairs with a similarity index ≥ 0.5 and a neutralization strength ≥ 5 (red area in **a**). **c**, Observed (dots) versus expected (violins) number of all $n = 303$ pairs with bNAb-imprinting capacity for various similarity thresholds (different subplots)

and neutralization strength thresholds (x axes). Violins (smoothed using a normal kernel) and one-sided P values were derived from 1,000 random reassignments of recipients to transmitters. Lines depict medians. **d**, Ratio of observed versus expected number of pairs with bNAb-imprinting capacity for various similarity thresholds (colours) and neutralization strength thresholds (x axis).

unlikely to be by chance. By randomly reassigning recipients to transmitters (10^6 replicates), we determined that the probability of such a strong shared neutralization strength by chance is $P_{\text{shuffle}} = 0.017$ (Fig. 2a). Although additional non-virus-linked factors that positively influence similar bNAb development may exist, our results point to a considerable influence of the transmitted virus in this elite bNAb-inducing pair. This strongly suggests the existence of virus envelopes with strong bNAb-imprinting capacity that will need to be identified for use in vaccine development.

We next developed a systematic, statistical approach (Extended Data Fig. 8) to identify pairs in which both partners developed notable neutralization strength (assessed by the lower neutralization score reached by one partner) and had similar neutralization and binding responses (Fig. 3a). Using various thresholds for the neutralization strength ($n = 7$) and the within-pair similarity index ($n = 4$) of the antibody response, we identified for each combination of thresholds the bNAb-imprinting capacity in our cohort. We counted the number of transmission pairs with a certain neutralization strength and a certain similarity index (exemplified for thresholds of neutralization = 5 and similarity = 0.5 in Fig. 3a, b). For all tested threshold criteria, we identified substantially more candidate transmission pairs with bNAb-imprinting capacity than would be expected by chance (Fig. 3c, d). This indicates that the transmitted HIV-1 strains in these pairs are promising candidates for imprinting highly similar neutralization responses.

Even though the viral Env antigen must certainly have a role, it is conceivable that bNAb-imprinting capacity might also, in part, depend on other virus traits, or demographic, host genomic, or transmission route-related factors. We therefore investigated the influence of known factors implicated in HIV-1 antibody development on bNAb-imprinting capacity in our cohort (Extended Data Fig. 7b, c). bNAb-imprinting capacity was detected more frequently among transmission pairs in which both partners had been infected for three or more years before assessment of antibody reactivity (Extended Data Fig. 7c). This is in line with the generally higher frequency of bNAb induction after prolonged exposure to HIV-1⁴. None of the other tested factors was significantly associated with bNAb-imprinting capacity (Extended Data Fig. 7b, c). Owing to the cross-sectional study set-up, we lacked the means to compare antibody responses at matched time points of infection and, therefore, our findings probably represent an underestimation of the frequency of bNAb-imprinters. Although our transmission cohort is one of the largest described to date, it may still require substantially larger cohorts and specifically tailored studies to detect more subtle influences of cofactors on bNAb imprinting.

The formal assessment of virus-induced heritability of the antibody response performed here quantifies the impact of HIV-1 immunogens on inducing qualitatively and quantitatively similar neutralizing and binding antibody responses. The effect of the infecting virus we observed across the assessed large, cross-sectional, natural history cohort of HIV-1 transmission pairs was significant, ranging between 7 and 19% depending on the antibody response considered (Fig. 1c). This is lower in magnitude than the effect of virus genetics on HIV-1 set point viral load (29%)^{7,8} but comparable to the effect of virus genetics on CD4⁺ T cell decline (17%)^{7,8} and the effect of host genetics of HLA and CCR5 SNPs (14.5%)⁹. Notably, owing to the cross-sectional nature of the transmission cohort, sampling time points of pairs were not matched for duration of infection. Adjustment for infection time controlled for this in the estimate of the overall imprinting capacity (Fig. 1c). Although in patients with shorter infection times, bNAb-imprinting capacity may have been missed (Fig. 3), the fact that we observe a significant genetic effect of the virus in a cross-sectional, natural history cohort is a clear indication of the strength of our results.

Using a statistical approach to systematically screen for virus strains with bNAb-imprinting capacity, we show that only a minority of HIV-1 transmissions exhibit transferability of strong neutralization antibody traits, suggesting that only some virus strains or Env variants harbour bNAb-imprinting capacity. bNAb-imprinting capacity may be a genuine

feature of certain Env variants or a temporal issue, as epitopes needed to evoke a specific neutralization response may be presented only transiently until escape mutations establish. Depending on when, with respect to viral evolution in the transmitter, the transmission occurred, the recipient may not have been exposed to the relevant epitope. If virus escape limits the heritability of antibody responses, transmission during acute HIV-1 infection (where neutralization activity is still low and hence neutralization escape is scarce) should show a higher similarity of antibody responses in transmitter and recipients. However, we found no evidence for higher frequencies of bNAb-imprinting capacity amongst acute phase transmission cases in our cohort (Extended Data Fig. 7c), suggesting that virus escape is not the main limiting factor and that higher patient numbers would be needed to uncover more subtle effects. In general, the fact that bNAb-imprinting upon transmission occurs strongly suggests that relevant epitopes that allow germline triggering can be preserved over longer time periods, for instance as a result of partial escape that retains bNAb-binding capacity^{10,11}.

Although the average effect of virus genetics across the entire cohort was moderate, individual cases with high antibody similarity and capacity to imprint a bNAb immune response can exist. The elite-neutralizing transmission pair that we identified highlights that distinct strains may exist, which harbour the potential to evoke highly similar bNAb and binding antibody responses across individuals. Further studies that identify more such pairs and a detailed characterization of the transmitted strains, longitudinal envelope evolution and the evoked bNAb responses will be needed to understand which specific features render a strain a bNAb imprinter. Clearly, if bNAb-imprinting strains exist, they need to be specifically investigated and characterized. In particular, envelope proteins from bNAb imprinters with proven in vivo transferability of antibody reactivity may provide the ultimate candidate immunogen(s) on which to base bNAb vaccine design.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0517-0>

Received: 12 February 2018; Accepted: 9 August 2018;

Published online: 10 September 2018

- Burton, D. R. & Hangartner, L. Broadly neutralizing antibodies to HIV and their role in vaccine design. *Annu. Rev. Immunol.* **34**, 635–659 (2016).
- Subbaraman, H., Schanz, M. & Trkola, A. Broadly neutralizing antibodies: what is needed to move from a rare event in HIV-1 infection to vaccine efficacy? *Retrovirology* **15**, 52 (2018).
- Mabuka, J., Goo, L., Omenda, M. M., Nduati, R. & Overbaugh, J. HIV-1 maternal and infant variants show similar sensitivity to broadly neutralizing antibodies, but sensitivity varies by subtype. *AIDS* **27**, 1535–1544 (2013).
- Rusert, P. et al. Determinants of HIV-1 broadly neutralizing antibody induction. *Nat. Med.* **22**, 1260–1267 (2016).
- Kadelka, C. et al. Distinct, IgG1-driven antibody response landscapes demarcate individuals with broadly HIV-1 neutralizing activity. *J. Exp. Med.* **215**, 1589–1608 (2018).
- Shankarappa, R. et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**, 10489–10502 (1999).
- Bertels, F. et al. Dissecting HIV virulence: heritability of setpoint viral load, CD4⁺ T cell decline and per-parasite pathogenicity. *Mol. Biol. Evol.* **35**, 27–37 (2018).
- Blanquart, F. et al. Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. *PLoS Biol.* **15**, e2001855 (2017).
- McLaren, P. J. et al. Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proc. Natl Acad. Sci. USA* **112**, 14658–14663 (2015).
- Bhiman, J. N. et al. Viral variants that initiate and drive maturation of V1V2-directed HIV-1 broadly neutralizing antibodies. *Nat. Med.* **21**, 1332–1336 (2015).
- Simonich, C. A. et al. HIV-1 neutralizing antibodies with limited hypermutation from an infant. *Cell* **166**, 77–87 (2016).

Acknowledgements Financial support for this study has been provided by the Swiss National Science Foundation (SNF; 314730_152663 and 314730_172790 to A.T.; 324730B_179571 to H.F.G.; PZ00P3-142411 and BSSG10_155851 to R.D.K.), the Clinical Priority Research Program of the University of Zurich (Viral infectious diseases: Zurich Primary HIV Infection Study to H.F.G. and A.T.), the Yvonne-Jacob Foundation (to H.F.G.), the Swiss

Vaccine Research Institute (to A.T., H.F.G. and R.D.K.) and the SystemsX.ch grant AntibodyX (to A.T.). This study has been cofinanced within the framework of the Swiss HIV Cohort Study, supported by the SNF (33CS30_148522 to H.F.G.), by the small nested SHCS project 744 (to A.T.) and by the SHCS research foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The SHCS data are collected by the five Swiss University Hospitals, two Cantonal Hospitals, 15 affiliated hospitals and 36 private physicians (listed in <http://www.shcs.ch/180-health-care-providers>). We thank the patients participating in the ZPHI and the SHCS and their physicians and study nurses for patient care and D. Perraudin and M. Minichiello for administrative assistance.

Reviewer information Nature thanks P. Lemey, J. Overbaugh and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.D.K., P.R., C.K., H.F.G. and A.T. conceived and designed the study and analysed data. M.Hu., A.M., H.E., M.S., T.L., N.F., J.W., T.U., N.S.B., C.L., H.K., J.B. and K.J.M. conducted experiments and analysed data. D.L.B., A.U.S., J.-P.C., M.C., E.B., M.Ho., A.C., M.B., A.R., S.Y., V.A., T.K., H.F.G. and the members of the Swiss HIV Cohort Study conceived and managed the SHCS and ZPHI cohorts, collected and contributed patient samples and clinical data. R.D.K., C.K., M.S., H.F.G. and A.T. wrote the manuscript, on which all co-authors commented.

Competing interests The University of Zurich filed a European patent application (EP18184854.0) that includes the full envelope sequences of patients T282 and R282 or components thereof for use as bNAb-inducing immunogens with R.D.K., P.R., H.F.G. and A.T. listed as inventors. All other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0517-0>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0517-0>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to R.D.K. or H.F.G. or A.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The Swiss HIV Cohort Study

Alexia Anagnostopoulos², Manuel Battegay⁸, Enos Bernasconi⁵, Jürg Böni¹, Dominique L. Braun², Heiner C. Bucher¹⁷, Alexandra Calmy⁷, Matthias Cavassini⁴, Angela Ciuffi¹⁸, Günter Dollenmaier¹⁹, Matthias Egger²⁰, Luigia Elzi³, Jan Fehr², Jacques Fellay²¹, Hansjakob Furrer³, Christoph A. Fux²², Huldrych F. Günthard², David Haerry²³, Barbara Hasse², Hans H. Hirsch^{8,12}, Matthias Hoffmann⁶, Irene Hösli²⁴, Michael Huber¹, Christian Kahlert^{6,25}, Laurent Kaiser¹⁰, Olivia Keiser²⁶, Thomas Klimkait¹², Roger D. Kouyos², Helen Kovari², Bruno Ledergerber², Gladys Martinetti²⁷, Begona Martinez de Tejada²⁸, Catia Marzolini⁸, Karin J. Metzner², Nicolas Müller², Dunja Nicca⁶, Paolo Paioni²⁹, Giuseppe Pantaleo¹¹, Matthieu Perreau¹¹, Andri Rauch⁹, Christoph Rudin³⁰, Alexandra U. Scherrer^{1,2}, Patrick Schmid⁶, Roberto Speck², Marcel Stöckle⁸, Philip Tarr³¹, Alexandra Trkola¹, Pietro Vernazza⁶, Gilles Wandeler⁹, Rainer Weber² & Sabine Yerly¹⁰

¹⁷Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, University of Basel, Basel, Switzerland.

¹⁸Institute of Microbiology, University Hospital Lausanne, University of Lausanne, Lausanne, Switzerland. ¹⁹Centre for Laboratory Medicine, Canton St. Gallen, St. Gallen, Switzerland. ²⁰Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland.

²¹Global Health Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ²²Clinic for Infectious Diseases and Hospital Hygiene, Kantonsspital Aarau, Aarau, Switzerland. ²³Positive Council, Zurich, Switzerland. ²⁴Clinic for Obstetrics, University Hospital Basel, University of Basel, Basel, Switzerland. ²⁵Children's Hospital of Eastern Switzerland, St. Gallen, Switzerland. ²⁶Institute of Global Health, University of Geneva, Geneva, Switzerland. ²⁷Cantonal Institute of Microbiology, Bellinzona, Bellinzona, Switzerland. ²⁸Department of Obstetrics and Gynecology, University Hospital Geneva, University of Geneva, Geneva, Switzerland. ²⁹University Children's Hospital, University of Zurich, Zurich, Switzerland. ³⁰University Children's Hospital, University of Basel, Basel, Switzerland. ³¹Kantonsspital Baselland, University of Basel, Basel, Switzerland

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were blinded to allocation (transmitter–recipient pairing) during initial antibody response analyses (all binding antibody responses and neutralization analyses to an eight-virus multi-clade panel. The investigators were not blinded during outcome assessment and extended neutralization analyses.

Study populations and ethics information. The starting cohort for our analysis were 4,281 chronic HIV-1-infected individuals included in a population-wide screen for HIV-1 neutralization breadth, now termed the Swiss 4.5K Screen⁴, for which extensive plasma neutralization and HIV-1 binding data were generated in a separate study⁵ (Extended Data Fig. 1a). Detailed information on the plasma sample/patient selection and study design of the Swiss 4.5K Screen has been previously described⁴ (Extended Data Fig. 1a). All plasma samples were collected during viraemic periods (no antiretroviral treatment at sampling time point) from adult, chronically infected HIV-1 patients. Because the cross-sectional sample selection was pre-determined by the Swiss 4.5K Screen, it was not possible during our current study to select samples matched for infection length or for closeness to putative transmission time points. Important for the interpretation of specific findings in the current study, the study design of the Swiss 4.5K Screen included three patient groups that differed in infection length (1–3, 3–5 and ≥ 5 years of untreated HIV-1 infection). The highest frequency of neutralization breadth was observed among individuals infected for > 3 years⁴. The plasma samples analysed previously^{4,5} and in the current study were provided by the biobanks of the Swiss HIV Cohort study (SHCS) and the Zurich Primary HIV Infection Study (ZPHI).

The SHCS and the ZPHI have been approved by the ethics committee of the participating institutions (Kantonale Ethikkommission Bern, Ethikkommission des Kantons St. Gallen, Comité départemental d'éthique des spécialités médicales et de médecine communautaire et de premier recours, Hôpitaux Cantonale de Genève, Kantonale Ethikkommission Zürich, Repubblica e Cantone Ticino—Comitato Etico Cantonale, Commission cantonale d'éthique de la recherche sur l'être humain, Canton de Vaud, Lausanne, Ethikkommission beider Basel for the SHCS and Kantonale Ethikkommission Zürich for the ZPHI) and written informed consent had been obtained from all participants. Please see the supplementary note of the previous study⁴ and the Supplementary Methods of the current study for further details on the cohorts and the collected patient data. The SHCS is a prospective, nationwide, longitudinal, non-interventional, observational, clinic-based cohort with semi-annual visits and blood collections, enrolling all HIV-infected adults living in Switzerland¹². Detailed information on the SHCS is available on <http://www.shcs.ch>. The SHCS, founded in 1988, is highly representative of the HIV epidemic in Switzerland as it includes an estimated 53% of all HIV cases diagnosed in Switzerland since the onset of the epidemic, 72% of all patients receiving antiretroviral treatment in Switzerland, and 69% of the nationwide-registered AIDS cases¹². The SHCS is registered under the Swiss National Science longitudinal platform: <http://www.snf.ch/en/funding/programmes/longitudinal-studies/Pages/default.aspx#Currently%20supported%20longitudinal%20studies>.

The ZPHI is a continuous, observational, non-randomized, single-centre cohort founded in 2002 that specifically enrolls patients with documented acute or recent primary HIV-1 infection (<https://www.clinicaltrials.gov/>; ID NCT00537966)¹³.

Establishment of the transmission pair cohort. Overview. We established a cohort of 303 putative transmission pairs by screening for potential transmission pairs within a cross-sectional starting cohort of 4,281 chronic HIV-1-infected individuals included in the Swiss 4.5K Screen^{4,5}. In addition to extensive patient demographic and clinical data, we had access to data on *pol* sequence, neutralization activity⁴ and HIV-1 binding antibody responses^{5,14} (Extended Data Fig. 1d) for these individuals.

We phylogenetically determined potential transmission pairs within the 4,281 patient starting cohort based on a threshold of *pol* gene similarity that was allowed to include individuals with a long infection history (Extended Data Fig. 1e, f and Supplementary Data 1). *pol* similarity has been previously established as reliable method to record genetic similarity of the infecting virus. Almost all larger molecular epidemiology work in HIV is based on HIV *pol*^{15–19}. Moreover, owing to genetic linkage, similarity in *pol* and similarity in *env* are highly correlated in transmission pairs^{13,20,21}. We determined potential transmission pairs as nearest neighbours (cherries) with a genetic distance of less than 0.045 on a *pol* phylogeny^{16,22} (Extended Data Fig. 1e, f and Supplementary Data 1). In total, we identified 303 potential transmission pairs and assigned transmitter and recipient status based on estimated infection dates. Individuals in transmission pairs were predominantly male, men who have sex with men and infected with subtype B (Supplementary Data 1 and Extended Data Table 1), reflecting the main drivers of domestic HIV-1 transmission in Switzerland^{4,12,13}. Overall, the cohort of 303 transmission pairs did not differ in terms of duration of infection, neutralization breadth, virus load and peripheral CD4 T cell counts from the remaining patients of the starting cohort (Extended Data Table 1). Specific steps conducted in the establishment of the cohort are described in full in the following sub-sections.

Construction of HIV-1 *pol* gene phylogenies. Potential transmission pairs were defined based on HIV-1 *pol* gene phylogenies (as described in detail previously^{16,22}) using *pol* nucleotide sequence data available in the SHCS database. In brief, these sequences stem from clinically or epidemiologically implicated genotypic resistance tests and, thus, are not derived from the same sampling time points as the plasma samples used for analyses of antibody responses.

We constructed *pol* phylogenies from 19,604 partial *pol* sequences from 10,970 different SHCS cohort participants, which included the 4,281 patients of the starting cohort (Extended Data Fig. 1d), and an additional 90,994 sequences from the Los Alamos database (<http://www.hiv.lanl.gov/>). These latter were included to decrease the chances of false-positive random clustering. We retrieved from the Los Alamos database all available, non-Swiss, *pol* sequences (region: 2,253–3,870) with a minimal length of 900 bp as of September 2014. Redundant control sequences (different sequence ID but identical nucleotide sequence) were deduplicated. For the SHCS patients, sequences with a minimum length of 250 bp for the protease gene and 500 bp for the RT gene were included. All sequences were initially aligned to a HXB2 reference genome (<http://www.ncbi.nlm.nih.gov/nucleotide/K03455.1>) using MUSCLE²³. Next, insertions relative to HXB2 and resistance mutations according to Stanford (<http://hivdb.stanford.edu/>) and International Antiviral Society—USA (<https://www.iasusa.org/>) lists were removed. In the following step, a generalized time-reversible model-based tree was constructed using FastTree²⁴. The R package 'APE' version 3.1 was used for tree exploration and analysis²⁵.

Definition of potential transmission pairs. Transmission pairs were defined as monophyletic pairs of the 4,281 starting cohort sequences on the *pol* phylogeny. We assumed that pairs with a cophenetic distance $> 4.5\%$, a commonly used threshold, clustered due to statistical and/or methodological artefacts (such as underrepresentation or even absence of rare genotypes in the background sequences). These pairs were therefore disregarded. We chose the relatively liberal threshold of pair distances of up to 4.5% as it best accommodates our cohort and the research questions addressed for several reasons. Owing to the study design of the Swiss 4.5K Screen, infection length varied between participants as one intent of this screen was to assess the influence of infection length on antibody development. Additionally, sampling time points for antibody testing and *pol* sequencing differ as the latter data were retrieved from clinically implicated genotypic resistance tests (see above). Short genetic distances in *pol* are associated with recent sampling times (Extended Data Fig. 1f). A strict distance criterion therefore bears the increased risk to exclude pairs in which plasma was sampled during prolonged chronic infection (at least 3 years of untreated infection), which we have previously shown to be associated with broad neutralization⁴. Thus, high genetic distances are a marker for the length of the within-patient evolution of the virus and associated with bNAb development, a main outcome of our study. A more liberal distance threshold therefore ascertains that the relevant patient population is included, maximizing the statistical power of our analysis. It is also important to note that, since we included a large number of background sequences, the detected monophyly is already a strong signal for the formation of a pair, irrespective of the actual genetic distance. We nevertheless verified the results of our study in sensitivity analyses using stricter selection criteria for potential transmission pairs, all of which confirmed the robustness of our observations (Extended Data Fig. 5).

Potential and confirmed transmission pairs. The attending physician confirmed that the bNAb transmission pair (T282 and R282) was a heterosexual couple living in a stable relationship in Switzerland, with a female-to-male transmission based on clinical data and self-reporting by the patients. Transmission of a HIV-1 subtype B infection from the female (North African origin) to the male partner (Asian origin) is further supported by a higher genetic *pol* diversity of the female partner (diversity = 0.53% for T282 and diversity = 0.20% for R282 on the same sampling date). On the basis of the available demographic data and infection timing, T282 infected R282 around 1.9 years post infection. The analysed plasma samples of the transmitter and recipient were collected 3.8 years and 3.4 years post infection, respectively. With the exception of T282 and R282, all other pairs should be viewed as potential transmission pairs as they possess genetically highly related virus strains in the *pol* phylogeny but have not been confirmed by alternative measures (physician records, self-reporting, and so on) and thus theoretically could also be partners in a tightly linked transmission cluster involving several people. For brevity we use the collective terms transmission pair(s)/transmission pair cohort without adding potential and/or confirmed to distinguish the respective patients. It is important to note that, as only the genetic relatedness of the infecting virus is of relevance for our study, neither the definition of confirmed pairs nor the tracing of the transmission time point is of relevance in the context of the conducted analyses.

Definition of transmitter and recipient within a pair. In a given transmission pair, we identified the individual with the earlier estimated infection date as the transmitter and the other as the recipient. In accordance with a previous study²⁶, our results were, however, not affected by this assumption. The analyses shown in Fig. 1c, Extended Data Figs. 3b, c, 5d, e and Extended Data Table 2 are by necessity symmetrical (that is, insensitive) with respect to which patient is identified

as transmitter and recipient. The analyses shown in Fig. 2a, b and Extended Data Figs. 4, 5a–c are not by necessity symmetrical, that is, they are potentially affected by which patient is identified as transmitter and recipient. We obtained, however, identical results when randomly assigning the role of transmitter and recipient within a transmission pair (Extended Data Fig. 3a). Note that this type of randomization does not change the grouping of patients into pairs but only which patient in a pair is considered transmitter or recipient.

HIV-1 binding antibody profile of the transmission pair cohort. The plasma IgG1, IgG2, IgG3 binding antibody reactivity to 13 HIV-1 Gag and Env antigens have been established for all 606 individuals in the 303 transmission pair cohort in a prior study analysing HIV-1 binding antibody of the 4,281 starting cohort⁵. See Supplementary Methods for details.

HIV-1 neutralizing antibody profile of the transmission pair cohort. Plasma neutralization activity of the 606 patients in the transmission pair cohort against a 14 multi-clade pseudovirus panel was determined as part of the current study. See Supplementary Methods for details.

Profiling of the neutralization breadth in the bNAb transmission pair T282 and R282. See Supplementary Methods.

Prediction of bNAb epitope specificity in plasma bNAb of transmission pair T282 and R282. See Supplementary Methods and a previously published study²⁷.

Strategies to determine the antibody-imprinting capacity of HIV-1. To investigate the similarity of antibody responses in transmission pairs, we used a series of approaches (summarized in Extended Data Fig. 2e and detailed below). In the context of our study, the strength of the overall effect of virus genetics on antibody responses measured by the above analyses is of secondary importance, as it can be the result of a strong influence of virus genetics in a few transmission cases and a weak/non-existent influence in most other cases within the cohort. Our analyses are thus tailored to provide proof of existence of an overall viral heritability, which—even if low—is a necessary condition for the search for bNAb imprinter viruses in a next step. The essence of what we investigate (see Fig. 1) is thus the similarity of the pattern of the responses across antigens/viruses and not that specific individuals have high antibody responses to all antigens/viruses.

Our analysis strategy followed three principle steps. First, we tested, for all 14 neutralization and 39 binding parameters, whether a given parameter is correlated within transmission pairs using Spearman correlations (ρ_{Spearman} for each parameter) (Fig. 1a and Extended Data Figs. 3a, 4a, c).

Second, the strength of the association is averaged across all neutralization, IgG1, IgG2 or IgG3 parameters to obtain a measure for the overall similarity of responses (average ρ_{Spearman} ; Fig. 1b and Extended Data Figs. 4b, d, 5a–c).

Third, a mixed-effect Tobit model was used to estimate the fraction of the variance explained by viral genetic factors, which can in addition be adjusted for the effect of covariables (Fig. 1c and Extended Data Figs. 3b, c, 5d, e). A Tobit model was used because the neutralization percentage data are always non-negative and the binding data are approximately uniformly distributed within the range of 0–1. The 303 pairs constitute the groups used in the mixed-effect Tobit model. The model estimates three types of variances, the group-level variance σ_P^2 , the individual level variance σ_I^2 , and (if cofactors are included in the model) the variance explained by the cofactors σ_C^2 . Accordingly, the heritability is then given by

$$h = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_I^2 + \sigma_C^2}$$

This constitutes a conservative approach to estimate heritability, since the variance explained by cofactors is also included in the denominator; that is, it is assumed to contribute to the non-heritable portion of the response although some cofactors (viral load, subtype and diversity) are at least partially steered by virus genetics. This approach is applied to each pseudovirus of the neutralization panel and each antigen separately and then averaged over all pseudoviruses/antigens in order to obtain an overall heritability estimate (Fig. 1c, Extended Data Figs. 3b, c, 5d, e and Extended Data Table 2).

In addition, we used linear mixed-effect models as an alternative method to derive heritability estimates, as these are commonly used^{26,28} and faster to be computed than Tobit models. The resulting estimates from variously adjusted linear mixed-effect models as well as standard errors derived from these models (by leave-one-out analyses) are compared to the Tobit estimates in Extended Data Fig. 3b. The standard errors proved to be small, and the linear mixed-effect models revealed almost identical effects to the Tobit mixed-effect models, highlighting the robustness of our findings.

In all three approaches, the statistical significance of a given heritability value (either for an individual pseudovirus or antigen, or averaged across pseudoviruses/antigens) is derived by comparing the observed value with the distribution obtained in 1,000 random reassignments of the recipients. The respective *P* value is then given by the inverse of the number of randomizations for which heritability is at least as high as in the original dataset. In Fig. 1b and Extended Data Figs. 4b, d,

5a–c, we consider two kinds of random reassignments: (1) completely random reassignment of recipients to transmitters and (2) random reassignment of recipients to transmitters with the same demographics (subtype (B versus non-B), ethnicity (white versus non-white), untreated infection length (1–3 years, 3–5 years, >5 years)). The latter adjusts for the potentially confounding effect of subtype, ethnicity and infection length. See also Supplementary Methods and previously published studies^{29,30} for details on how we controlled for potential confounders.

Single-genome amplification of the HIV-1 envelope. See Supplementary Methods and a previous study³¹.

HIV-1 envelope cloning from undiluted cDNA. See Supplementary Methods.

Infusion vector cloning and sequencing. See Supplementary Methods.

HIV-1 full-length genome sequencing and analysis. See Supplementary Methods and previously published studies^{32–36}.

HIV-1 envelope phylogenetic analysis. See Supplementary Methods and previously published studies^{37–39}.

Systematic strategy to trace HIV-1 strains with bNAb-imprinting capacity. To derive a method that allows a systematic identification of HIV-1 strains that have the capacity to induce (imprint) bNAb activity, we used a multi-step approach summarized in Extended Data Fig. 8.

Step 1. For each transmission pair, we test whether the antibody response across pseudo-viruses or antigens is correlated. This indicates whether the two members of a pair have high responses against the same viruses or antigens. Specifically, we determine the Spearman correlation coefficient of the neutralization responses in transmitter and recipient, and the same is done for the IgG1-binding response for each transmission pair. The two correlation coefficients are combined to obtain a single antibody response similarity index ($\frac{1}{2}\rho_{\text{Spearman-neutralization}} + \frac{1}{2}\rho_{\text{Spearman-IgG1-binding}}$). Note, we only use IgG1 binding responses as these dominate natural HIV-1 infections and also provided the highest within-pair similarity (average $\rho_{\text{Spearman}} = 0.19$, $P_{\text{shuffling}} < 0.001$; Fig. 1a, b).

Step 2. Because a large fraction of patients exhibits weak responses against all tested pseudoviruses (148 out of 606 patients distributed over 127 pairs neutralize each of the 14 viruses at <20%, that is, neutralization score 0), and the comparison of their fingerprints is thus uninformative, this fingerprint-based approach had to be restricted to pairs in which both patients exhibit a certain level of neutralization strength (as measured by the neutralization score). For each transmission pair, we calculated the neutralization strength as the minimum neutralization score of the two individuals.

Step 3. Transmission pairs with highly similar antibody binding responses and high neutralization scores can now be identified (Fig. 3a, b).

Step 4. The number of transmission pairs with a similar (binding) and strong (neutralization) antibody response is determined for a broad range of similarity (0.3, 0.4, 0.5 and 0.6) and strength thresholds (1, 2, ..., 7). To determine significance, these numbers were compared to those observed in 1,000 replicate datasets with shuffled transmission pairs, for which the shuffling is realized by assigning a randomly chosen recipient (sampling without replacement) to each transmitter (Fig. 3c, d).

For the interpretation of the results, it is important to note that not all pairs can be expected to show similar responses as in particular neutralization breadth is genuinely rare. In addition, in cross-sectional analyses as ours, it must be anticipated that occasionally the effects of the infecting virus are masked by the effect of differential infection length or other confounding factors in transmitter and recipient. We therefore control intensely for confounding factors (see also section ‘Strategies to determine the antibody-imprinting capacity of HIV-1’). The clear overall shift towards positively correlated responses (Fig. 3a, b), which is not a chance finding as we show (Fig. 3c, d), highlights that a considerable fraction of viruses exists that induce similar responses. Among these, rare cases with bNAb imprinting capacity are expected and can be screened for.

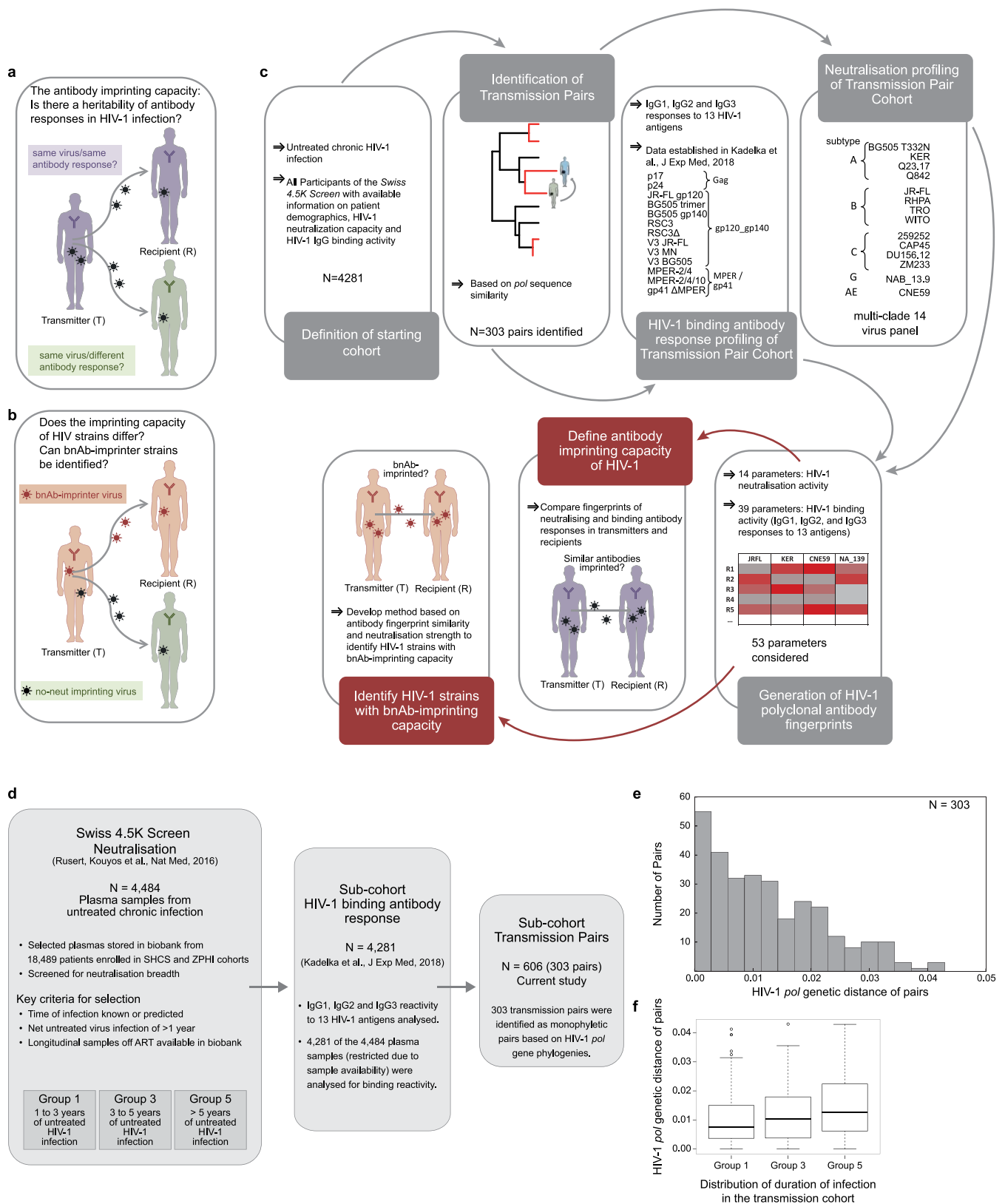
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The antibody response and patient data reported in this paper are completely tabulated in Supplementary Data 1, 2. Sequence data of T282–R282 Env consensus and Env clones reported in Fig. 2c are deposited in GenBank. Accession codes are listed in Extended Data Fig. 6b. The raw sequencing files of the Illumina full HIV sequencing data of patients T282 and R282 referred to in Fig. 2c and Extended Data Fig. 6 have been uploaded to <https://zenodo.org/> (<https://doi.org/10.5281/zenodo.1324259>). *pol* sequence data of the 606 studied cases are available from the corresponding authors and/or the SHCS scientific board (<http://www.shcs.ch/contact>) upon request.

12. Schoeni-Affolter, F. et al. Cohort profile: the Swiss HIV Cohort study. *Int. J. Epidemiol.* **39**, 1179–1189 (2010).

13. Rieder, P. et al. Characterization of human immunodeficiency virus type 1 (HIV-1) diversity and tropism in 145 patients with primary HIV-1 infection. *Clin. Infect. Dis.* **53**, 1271–1279 (2011).

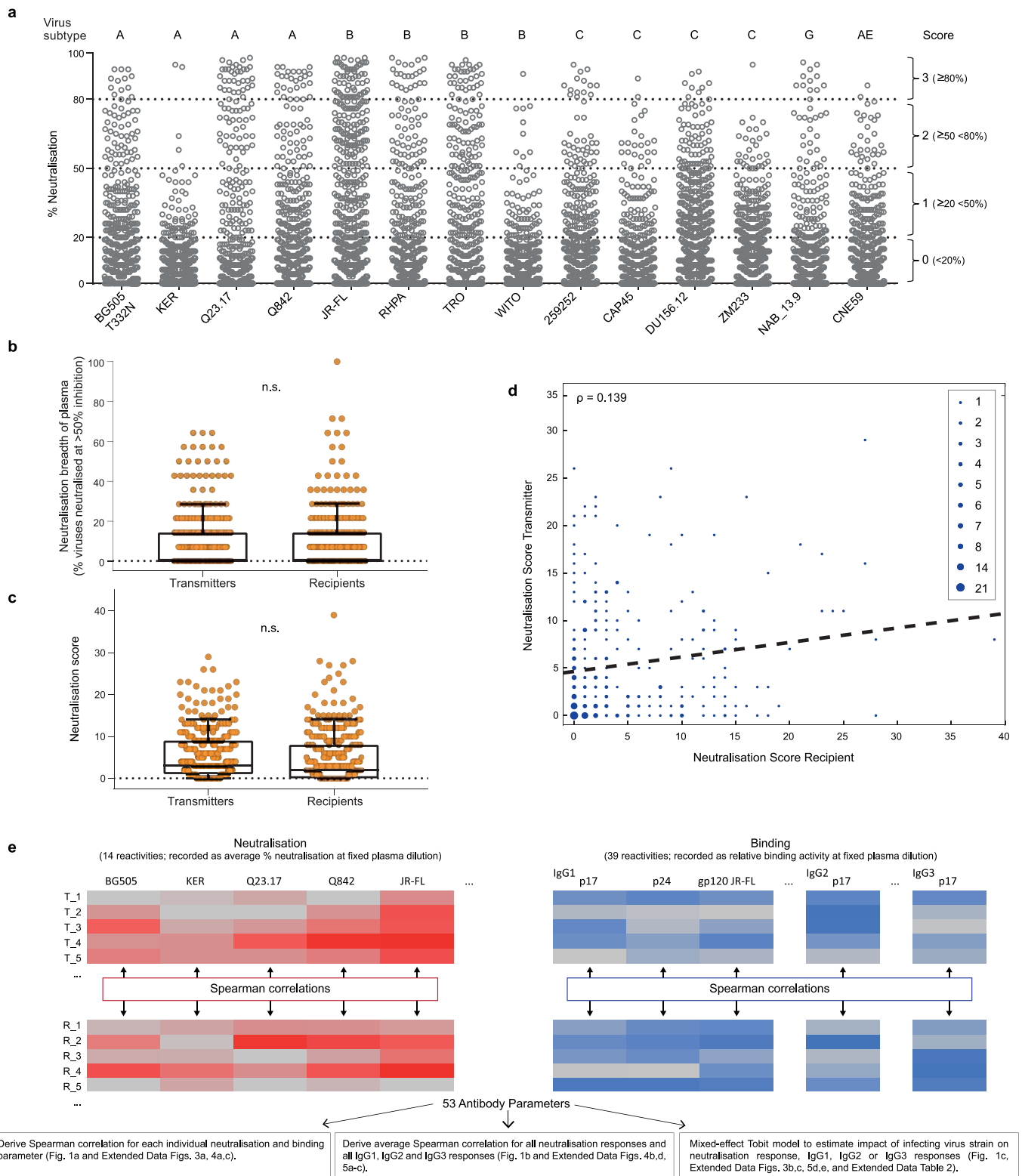
14. Liechti, T. et al. Development of a high-throughput bead based assay system to measure HIV-1 specific immune signatures in clinical samples. *J. Immunol. Methods* **454**, 48–58 (2018).
15. von Wyl, V. et al. The role of migration and domestic transmission in the spread of HIV-1 non-B subtypes in Switzerland. *J. Infect. Dis.* **204**, 1095–1103 (2011).
16. Kouyos, R. D. et al. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J. Infect. Dis.* **201**, 1488–1497 (2010).
17. Brenner, B., Wainberg, M. A. & Roger, M. Phylogenetic inferences on HIV-1 transmission: implications for the design of prevention and treatment interventions. *AIDS* **27**, 1045–1057 (2013).
18. Wertheim, J. O. et al. The global transmission network of HIV-1. *J. Infect. Dis.* **209**, 304–313 (2014).
19. Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J. & Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data. *AIDS* **31**, 1211–1222 (2017).
20. Hué, S., Clewley, J. P., Cane, P. A. & Pillay, D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* **18**, 719–728 (2004).
21. Oberle, C. S. et al. Tracing HIV-1 transmission: envelope traits of HIV-1 transmitter and recipient pairs. *Retrovirology* **13**, 62 (2016).
22. Marzel, A. et al. HIV-1 transmission during recent infection and during treatment interruptions as major drivers of new infections in the Swiss HIV Cohort Study. *Clin. Infect. Dis.* **62**, 115–122 (2016).
23. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
24. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
25. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
26. Bachmann, N. et al. Parent–offspring regression to estimate the heritability of an HIV-1 trait in a realistic setup. *Retrovirology* **14**, 33 (2017).
27. Tiller, T. et al. Efficient generation of monoclonal antibodies from single human B cells by single cell RT–PCR and expression vector cloning. *J. Immunol. Methods* **329**, 112–124 (2008).
28. Mitov, V. & Stadler, T. A practical guide to estimating the heritability of pathogen traits. *Mol. Biol. Evol.* **35**, 756–772 (2018).
29. Venner, C. M. et al. Infecting HIV-1 subtype predicts disease progression in women of sub-Saharan Africa. *EBioMedicine* **13**, 305–314 (2016).
30. Alizon, S. & Fraser, C. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology* **10**, 49 (2013).
31. Salazar-Gonzalez, J. F. et al. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J. Virol.* **82**, 3952–3970 (2008).
32. Giallonardo, F. D. et al. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic Acids Res.* **42**, e115 (2014).
33. Gall, A. et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. *J. Clin. Microbiol.* **50**, 3838–3844 (2012).
34. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
35. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at <https://arxiv.org/abs/1207.3907> (2012).
36. Wilm, A. et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).
37. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
38. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
39. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).



Extended Data Fig. 1 | See next page for caption.

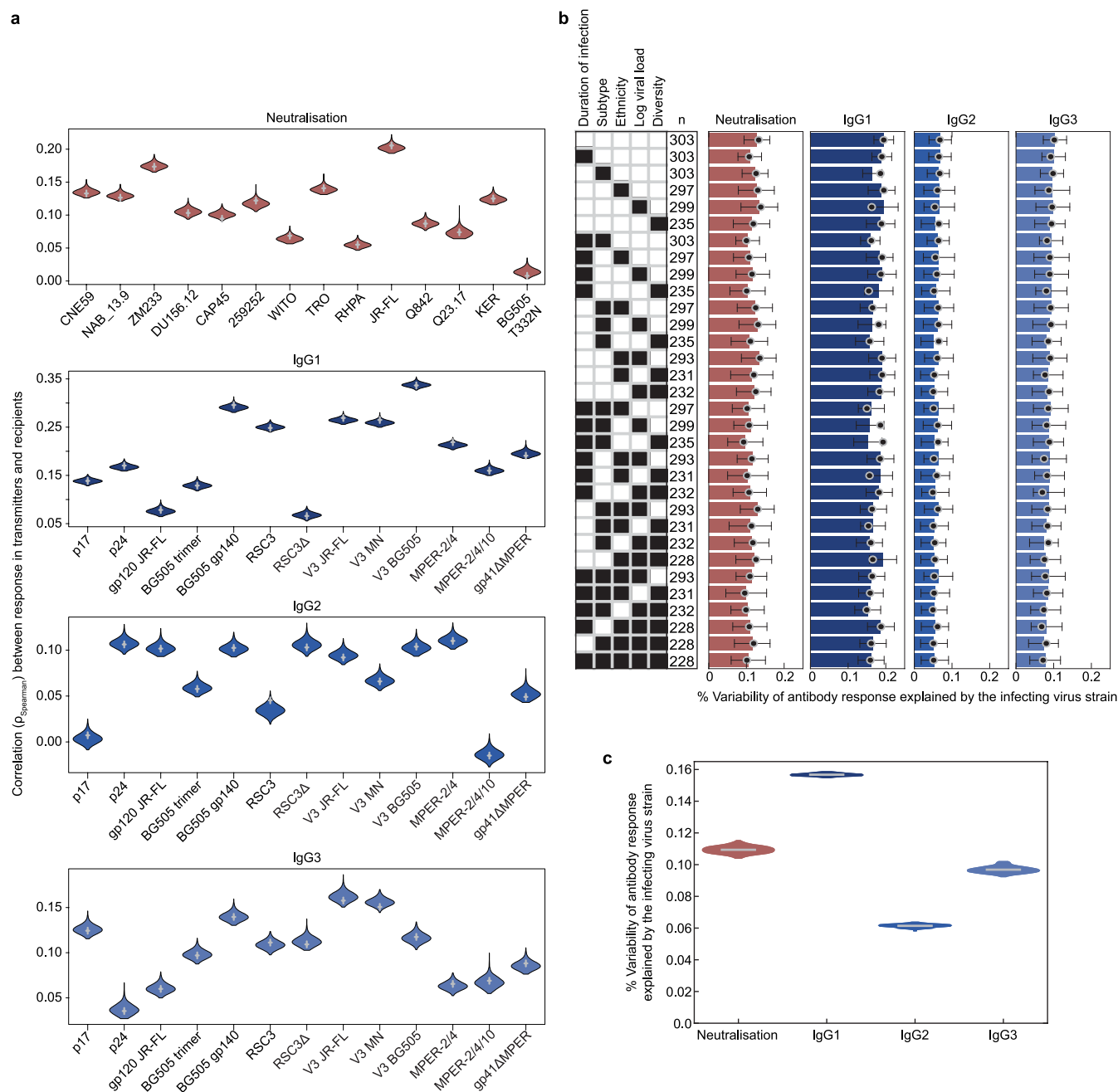
Extended Data Fig. 1 | Selection and characterization of the HIV-1 transmission cohort used to investigate the antibody imprinting capacity of HIV-1. a, b, Principal aims of the study. **a,** Defining the binding and neutralizing antibody imprinting capacity of HIV-1 strains/immunogens (that is, determining the heritability of these responses). The schematic depicts the possible outcomes of a transmission. The recipient may either have the same or a different type of antibody response as the transmitter. **b,** Creating means to identify bNAb-imprinting HIV-1 strains. The schematic depicts the possible outcomes of a transmission if the transmitter has a bNAb response. The recipient may also develop a bNAb response or not. **c,** Workflow of the study. **d,** The search for transmission

pairs started with a cohort of 4,281 patients, who were included in the Swiss 4.5K Screen⁴ and for which HIV-1 antibody-binding data are available⁵. **e,** Histogram of the genetic (*pol*) distance of the 303 identified transmission pairs. **f,** Genetic (*pol*) distance to the transmission partner stratified by duration of untreated infection for the $n = 606$ patients in the transmission pair cohort. As in the Swiss 4.5K Screen, patients are grouped by infection length into categories of 1–3 years ($n = 138$), 3–5 years ($n = 252$) and more than 5 years ($n = 216$). Medians are shown (centre line), each box spans the IQR and each whisker extends to the most extreme value no more than $1.5 \times$ IQR from the box. More extreme values are shown as points.



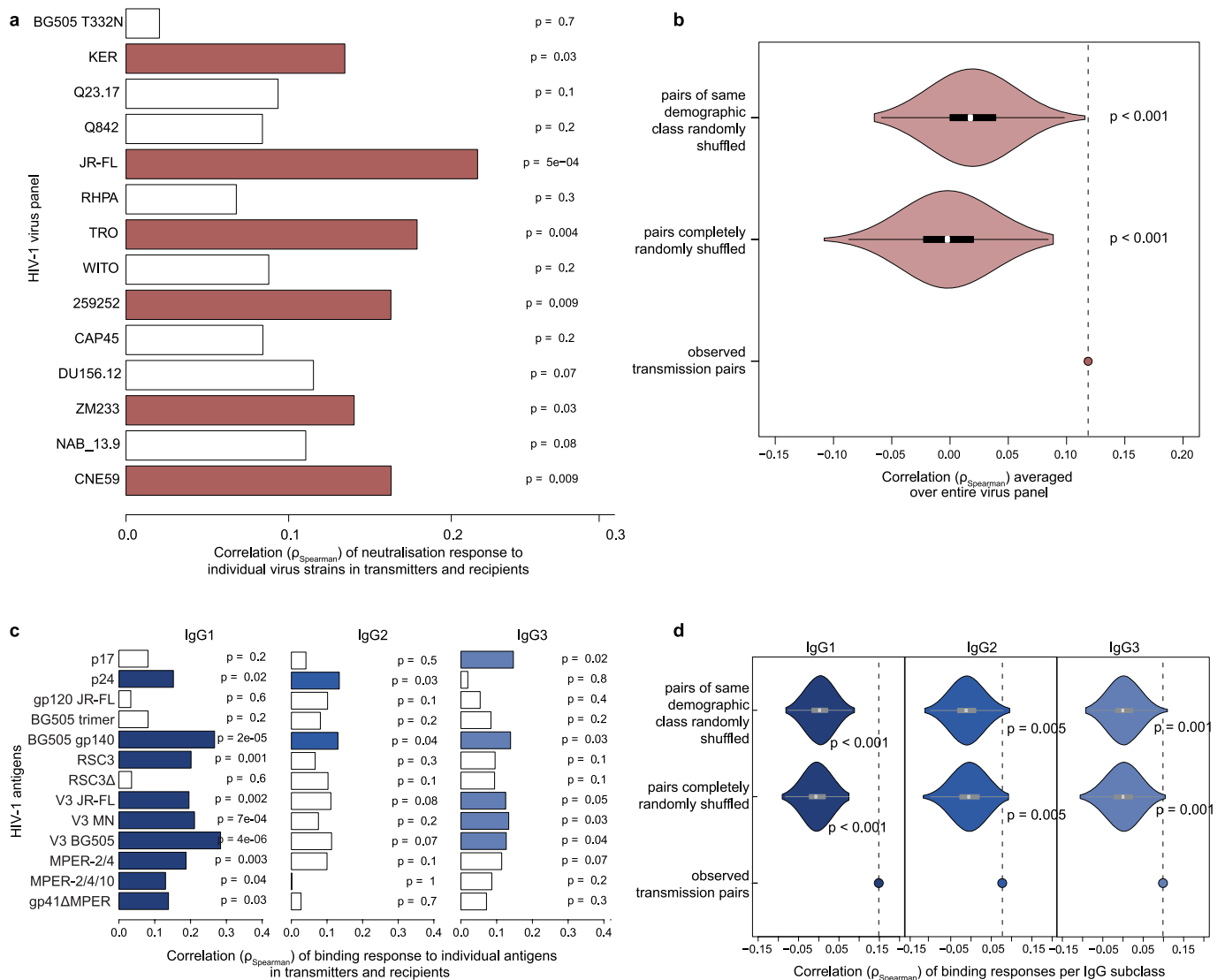
Extended Data Fig. 2 | Neutralization score analyses and strategies to determine the antibody-imprinting capacity of HIV-1. **a**, Neutralization activity against a multi-clade 14-virus panel of all 606 transmitter and recipient plasma samples. Dashed horizontal lines at 20%, 50% and 80% and brackets display the thresholds used for the assignment of scores. On the basis of these data, each plasma–virus combination received a score of 0–3. The neutralization score of a plasma sample is the sum of the 14 individual scores against the panel viruses (0–42). Individual data points are shown as jittered circles. **b**, **c**, Distribution of neutralization breadth (**b**; $P = 0.69$) and neutralization scores (**c**; $P = 0.40$) in $n = 303$ transmitters and recipients (P values: two-tailed Wilcoxon signed-rank test). Medians

are shown (centre line), each box spans the IQR and the whiskers extend to the 10% and 90% percentile. Individual data points are shown as jittered circles. **d**, Scatter plot of the neutralization score of recipients and transmitters for all 303 transmission pairs (Spearman; $\rho = 0.139$, $P = 0.015$). The number of pairs with the indicated values is depicted by dot size. **e**, Strategies to determine the antibody-imprinting capacity of HIV-1. For each of the 14 neutralization and 39 binding reactivities, the similarity in reactivity (Spearman correlation) was compared between transmitters and recipients to test whether transmission partners shared similar antibody responses.



Extended Data Fig. 3 | Antibody similarity measurements are insensitive to the assignment of transmitter and recipient status, to choice of confounders and outliers. **a**, Distribution of Spearman correlations (as in Fig. 1a) of neutralization–antibody-binding responses in pairs, in which the role of transmitter and recipient was randomly assigned within each of the $n = 303$ pairs (1,000 reassignments). Each violin plot is smoothed using a normal kernel, and its width represents the likelihood of a certain Spearman correlation. Grey daggers correspond to the similarities retrieved for the actual assignment of transmitter and recipient used in this study. For all investigated parameters, there was no statistically significant difference between the actual assignment and the shuffled assignment of transmitters and recipients (two-sided $P_{\text{Spearman}} > 0.05$). **b**, Proportion of variability in responses explained

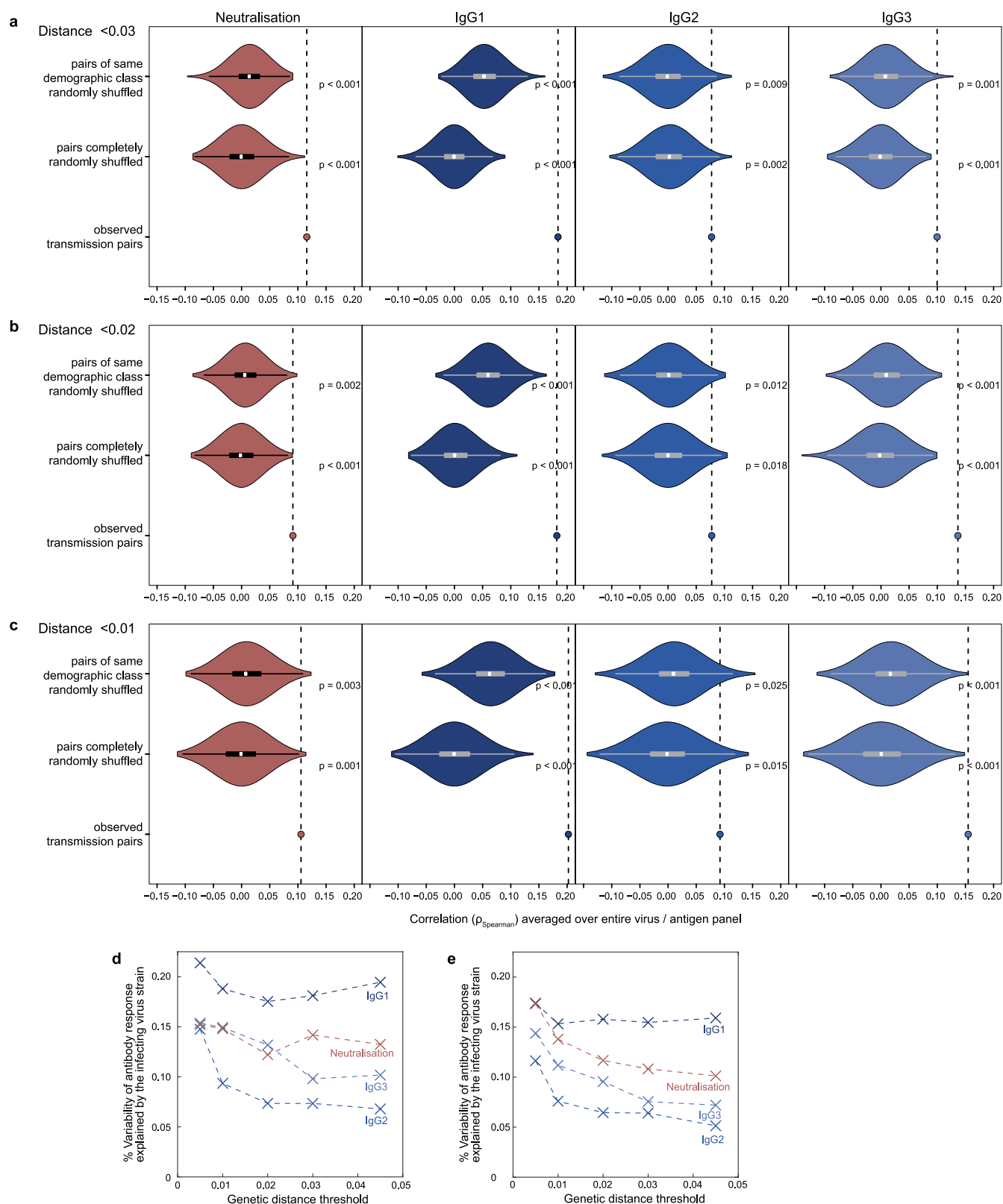
by the infecting virus, determined using mixed-effect Tobit regression models (black circles) and linear mixed-effect models (horizontal bars). Standard errors of the linear mixed-effect models (black error bar; obtained by leave-one-out analyses) are shown. In each row, the models were adjusted for the indicated set of cofactors and only transmission pairs with complete information were included (number of pairs shown per row). **c**, Proportion of variability in responses explained by the infecting virus, determined using mixed-effect Tobit regression models on the full 303 pairs for which missing cofactors were imputed based on the other cofactors. Each violin plot is the result of 100 independent imputations, smoothed using a normal kernel, and its width represents the likelihood of a certain variability. Medians are shown (lines).



Extended Data Fig. 4 | Similarity of neutralization and antibody-binding responses in subtype-B-infected transmission pairs.

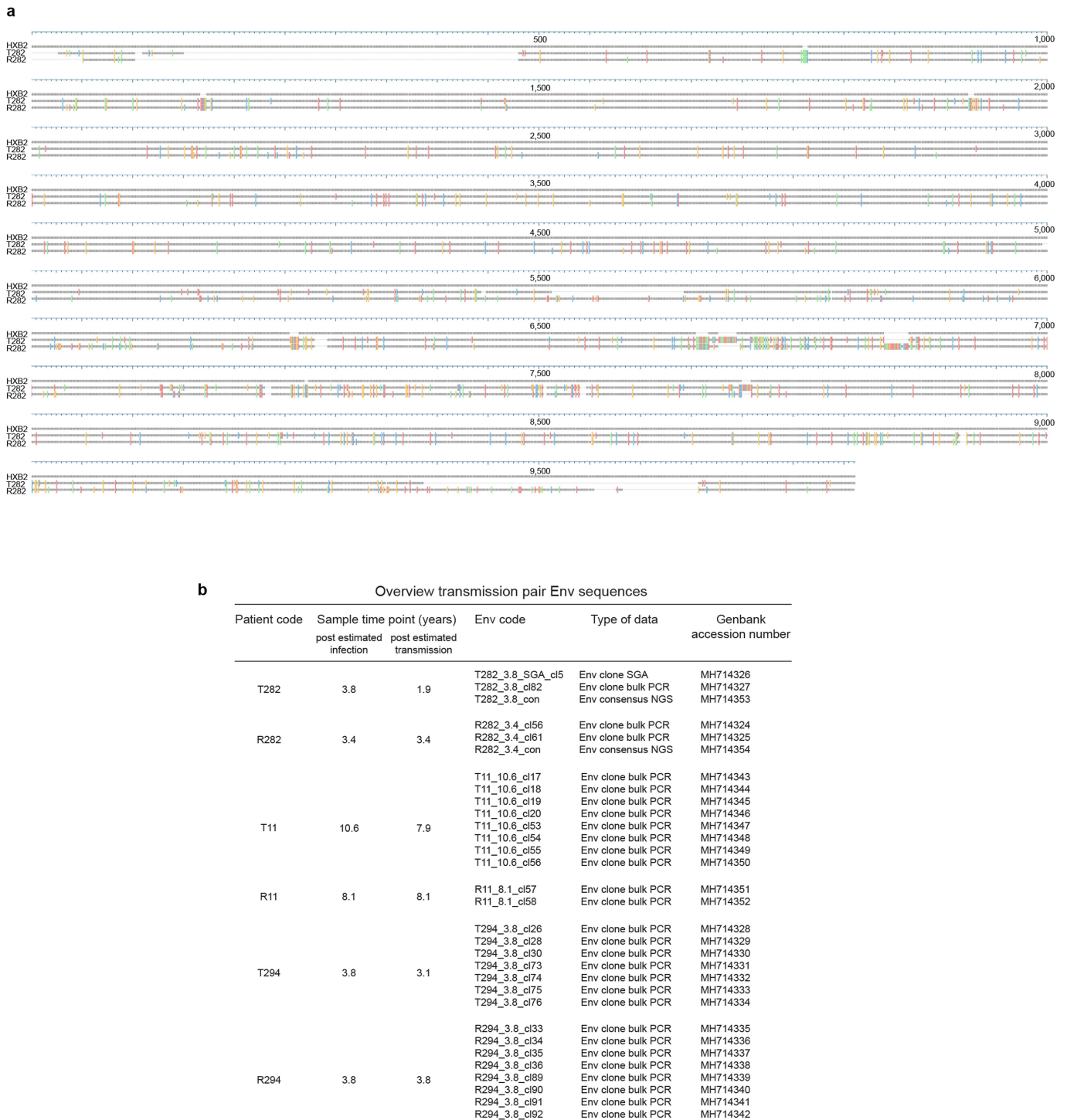
a–d, To exclude influences of the subtype of the infecting virus, as a sensitivity analysis to Fig. 1a, b, the similarity of neutralization (**a**, **b**) and antibody-binding (**c**, **d**) responses was tested for the subset of subtype-B-infected transmission pairs ($n = 254$). **a**, **c**, Spearman correlation of the neutralization–antibody-binding response to each pseudovirus or antigen. Significant correlations (two-sided $P_{\text{Spearman}} < 0.05$) are coloured. **b**, **d**, Average Spearman correlation of antibody responses in observed transmission pairs ($n = 303$) compared to two alternative scenarios:

(1) completely random reassignment of recipients to transmitters and (2) random reassignment of recipients to transmitters with same demographics (subtype, ethnicity and untreated infection length). One-sided P values were derived from comparison with 1,000 reassignments. Each violin plot is smoothed using a normal kernel, and its width represents the likelihood of a certain average correlation in the respective alternative scenario. The medians are shown (white dots), each box spans the IQR and each whisker extends to the most extreme value no more than $1.5 \times \text{IQR}$ from the box.



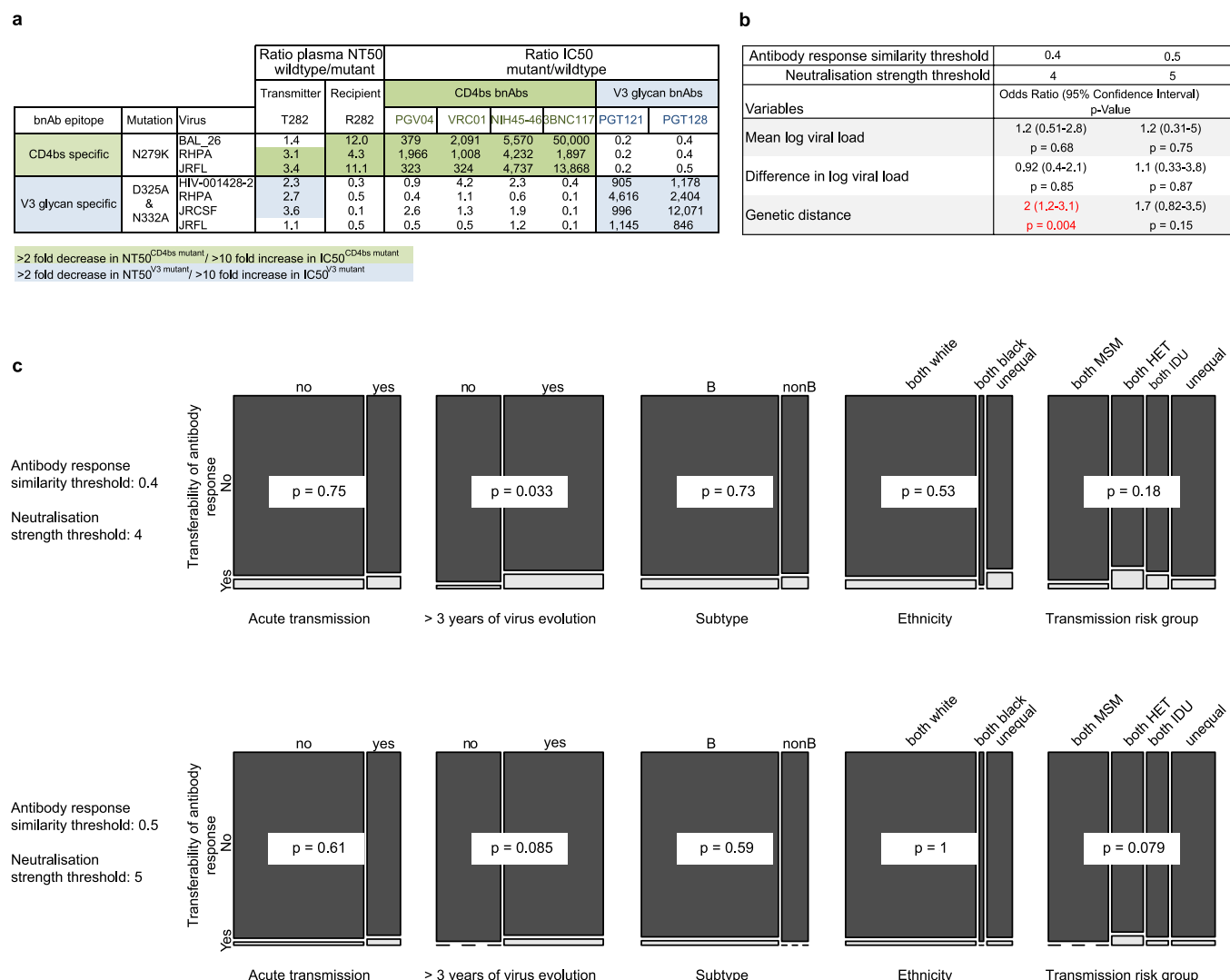
Extended Data Fig. 5 | Within-pair similarity of neutralization and binding responses remains significant for transmission pairs with lower genetic distance. a–c. As a sensitivity analysis for Fig. 1b, the average similarity of neutralization and antibody-binding responses was determined for those transmission pairs with a genetic distance <0.03 (a, $n = 280$), <0.02 (b, $n = 243$) or <0.01 (c, $n = 148$), and compared to two alternative scenarios: (1) completely random reassignment of recipients to transmitters and (2) random reassignment of recipients to transmitters with same demographics (subtype, ethnicity and untreated infection length). One-sided P values were derived from comparison with

1,000 reassignments. Each violin plot is smoothed using a normal kernel, and its width represents the likelihood of a certain average correlation in the respective alternative scenario. The medians are shown (white dots), each box spans the IQR and each whisker extends to the most extreme value no more than $1.5 \times$ IQR from the box. **d, e.** As a sensitivity analysis for Fig. 1c, the proportion of variability in responses explained by the infecting virus, determined using unadjusted (**d**) and fully adjusted (**e**) (adjusted for duration of infection, subtype, ethnicity, log viral load and diversity) mixed-effect Tobit regression models is shown when restricting the analysis to closely related pairs (threshold on the x axis).



Extended Data Fig. 6 | Full-genome comparison of T282 and R282 consensus sequences. a, Multiple sequence alignment of the full genome consensus sequences of T282 and R282 with HXB2. Nucleotide variations from HXB2 are depicted by colour: A (red), T (green), C (blue),

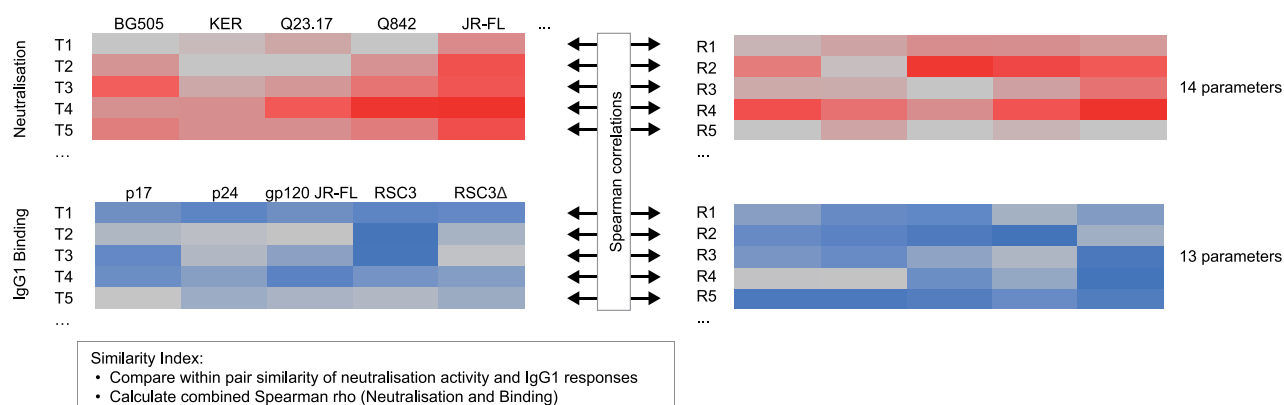
G (yellow); deletions by a horizontal bar. **b,** Overview and accession codes of *env* sequences derived from three bNAb-imprinting transmission pairs depicted in Fig. 2c. Indicated transmission time estimate is based on estimated time of infection of recipient.



Extended Data Fig. 7 | Mutant virus neutralization mapping and distribution of transmission pair characteristics among pairs with and without transferability of antibody response. a, Changes in neutralization activity of wild-type and mutant viruses were compared for six bNAbs (four targeting the CD4 binding site, two targeting the V3 glycan) and for the plasma of the elite-neutralizing pair. Increases in mutant half-maximum inhibitory concentration (IC₅₀) values >tenfold and decreases in mutant half-maximum neutralization titre (NT₅₀) values >two fold appear coloured. **b, c,** Association of transmission pair characteristics and transferability of antibody response in *n* = 303 transmission pairs. Transferability of antibody responses was determined

according to relatively liberal (strength threshold = 4 and similarity threshold = 0.4) and strict criteria (strength threshold = 5 and similarity threshold = 0.5). **b,** Influence of continuous variables (mean log₁₀ HIV-1 RNA in the transmission pair, the difference of these log₁₀ RNA values, and the genetic distance in the pair) tested by univariable logistic regression. **c,** Influence of categorical variables (acute infection, long virus evolution, infecting subtype, ethnicity and transmission mode) tested by two-tailed Fisher's exact test. The area of a rectangle corresponds to the number of pairs with the respective characteristic. MSM, men who have sex with men; HET, heterosexual transmission; IDU, intravenous drug users.

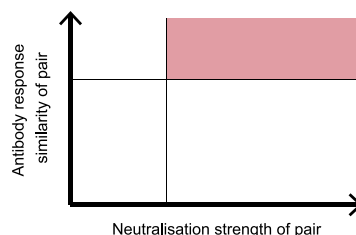
Step 1: Establish antibody response similarity index



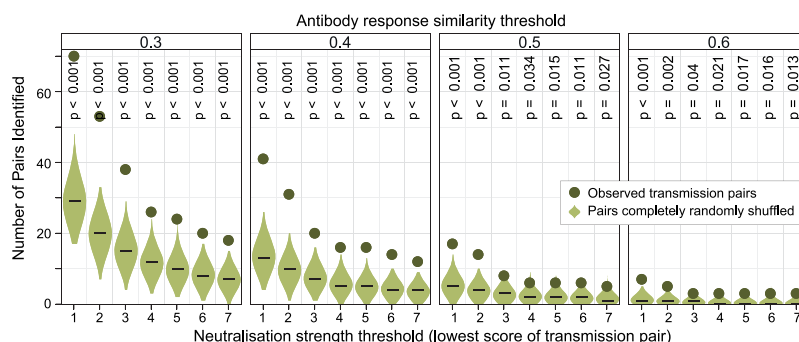
Step 2: Establish neutralisation strength of pair

- calculate lowest neutralisation score of pair

Step 3: Determine pairs with high similarity of antibody response and high neutralisation score



Step 4: Probe thresholds of similarity index and high neutralisation score to define HIV-1 strains with higher than random capacity to imprint neutralising antibody responses



Extended Data Fig. 8 | Systematic strategy to trace HIV-1 strains with bNAb-imprinting capacity. HIV-1 strains that have the capacity to induce bNAb activity can be identified by measuring the similarity and the

strength of the antibody response in each pair. A stepwise description of the approach is shown. See Methods ‘Systematic strategy to trace HIV-1 strains with bNAb-imprinting capacity’ for details.

Extended Data Table 1 | Patient demographics of the study population and entire population in the Swiss 4.5K screen

	Not part of a transmission pair	Part of a transmission pair
n	3675	606
<i>pol</i> Subtype (%)		
A	343 (9.3)	25 (4.1)
B	2514 (68.4)	510 (84.2)
C	138 (3.8)	8 (1.3)
G	51 (1.4)	6 (1.0)
AE	135 (3.7)	16 (2.6)
Other ^a	266 (7.2)	41 (6.8)
Unknown	228 (6.2)	0 (0.0)
Untreated Infection Length (%)		
1-3 years	701 (19.1)	138 (22.8)
3-5 years	1445 (39.3)	252 (41.6)
5 or more years	1529 (41.6)	216 (35.6)
Ethnicity (%)		
White	2878 (78.3)	534 (88.1)
Black	509 (13.9)	29 (4.8)
Other ^b	246 (6.7)	36 (5.9)
Unknown	42 (1.1)	7 (1.2)
Transmission Mode ^c (%)		
Hetero	1370 (37.3)	188 (31.0)
MSM	1452 (39.5)	275 (45.4)
IDU	726 (19.8)	127 (21.0)
Other	127 (3.5)	16 (2.6)
bnAb activity (%)		
None or weak	2868 (78.0)	493 (81.4)
Cross	604 (16.4)	80 (13.2)
Broad	152 (4.1)	26 (4.3)
Elite	51 (1.4)	7 (1.2)
Median Viral Diversity ^d [IQR]	0.00902 [0.00317, 0.01767]	0.00668 [0.00154, 0.01466]
Median Viral Load (copies/ml; Log ₁₀) [IQR]	4.237 [3.608, 4.747]	4.317 [3.771, 4.830]
Median CD4 Count (cells/ μ l) [IQR]	407.0 [305.0, 550.0]	415.5 [299.75, 561.0]

^aThe subtype 'Other' includes all patients infected with low-frequency subtypes and circulating recombinant forms.

^bThe ethnicity 'Other' includes mostly patients with Asian and Hispanic origin.

^cMSM, men who had sex with men; IDU, intravenous drug users; Hetero, heterosexual transmission; 'Other' includes all patients who acquired their HIV infection by blood products, prenatally, or with unknown transmission mode.

^dViral diversity was measured as *pol* gene sequence ambiguity.

Extended Data Table 2 | Proportion of variability of neutralization and binding responses explained by the infecting virus strain using variously adjusted mixed-effect Tobit models

Possible confounders included		Unadjusted models				Adjusted models	
Duration of infection & subtype & ethnicity	No	Yes	Yes	Yes	Yes	Yes	Yes
Log Viral Load	No	No	Yes	Yes	No	Yes	Yes
Viral Diversity	No	No	No	Yes	No	No	Yes
Sample size	303*2	297*2	293*2	228*2	297*2	293*2	228*2
% of the variability of the antibody response explained by virus factors							
Pseudovirus / Antigen		p-value					
Neutralisation	BG505N332	2.8	2.4	2.6	6.4	1.3	2.2
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	KER	21.1	21.1	21.0	14.5	22.1	22.3
		p < 0.001	p < 0.001	p < 0.001	p = 0.022	p = 0.001	p = 0.018
	Q23	10.2	11.2	10.9	8.8	9.9	8.8
		p = 0.047	p = 0.04	p = 0.042	n.s.	n.s.	n.s.
	Q842	13.4	12.9	13.1	14.7	10.7	12.6
		p = 0.023	p = 0.021	p = 0.02	p = 0.026	p = 0.049	p = 0.035
	JRFL	23.0	21.9	23.5	15.4	16.7	18.9
		p < 0.001	p < 0.001	p < 0.001	p = 0.015	p = 0.005	p = 0.004
	RHPA	5.0	4.9	5.1	4.1	2.9	4.1
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	TRO	17.5	16.3	16.3	10.8	13.9	15.2
		p = 0.004	p = 0.009	p = 0.008	n.s.	p = 0.02	p = 0.015
	WITO	14.4	13.1	13.9	11.3	14.0	14.8
IgG1 binding		p = 0.01	p = 0.025	p = 0.015	n.s.	p = 0.016	p = 0.017
	V25	10.1	10.5	10.6	7.1	8.2	9.7
		p = 0.045	p = 0.039	p = 0.04	n.s.	n.s.	n.s.
	CAP45	16.9	17.0	16.9	14.6	12.0	12.6
		p = 0.002	p = 0.003	p = 0.003	p = 0.023	p = 0.029	p = 0.022
	DU156	11.7	12.5	12.5	10.8	10.9	11.9
		p = 0.021	p = 0.015	p = 0.016	n.s.	p = 0.043	p = 0.03
	ZM233	17.8	17.6	17.5	12.0	13.1	14.2
		p = 0.003	p = 0.001	p = 0.001	p = 0.033	p = 0.02	p = 0.015
	NAB_139	10.1	9.9	10.0	6.3	10.8	11.1
		n.s.	n.s.	n.s.	n.s.	n.s.	p = 0.05
	CNE59	11.4	11.6	11.5	12.3	6.7	6.7
		p = 0.048	p = 0.035	p = 0.039	p = 0.047	n.s.	n.s.
	p17	14.6	14.5	14.5	10.4	9.3	8.0
		p = 0.005	p = 0.008	p = 0.007	p = 0.03	p = 0.03	p = 0.05
IgG2 binding	p24	17.2	17.9	17.6	15.0	14.0	12.3
		p = 0.002	p = 0.001	p = 0.001	p = 0.018	p = 0.004	p = 0.016
	JR-FL gp120	7.3	5.8	7.1	10.5	2.0	3.5
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	BG505 trimer	12.6	13.3	14.1	11.7	12.0	13.9
		p = 0.011	p = 0.009	p = 0.006	p = 0.045	p = 0.016	p = 0.01
	BG505 gp140	27.7	28.5	29.6	31.0	26.6	28.5
		p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001
	RSC3	24.8	24.2	25.0	25.7	17.7	18.4
		p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001
	RSC3Δ	6.8	5.7	6.2	0.0	4.7	6.1
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	V3 JR-FL	26.5	28.1	28.9	25.1	25.0	29.6
		p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001
	V3 MN	25.8	27.6	27.6	31.4	26.1	26.0
IgG3 binding		p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001
	V3 BG505	32.0	33.5	33.3	32.1	28.1	28.1
		p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001
	MPER-2/4	21.1	21.4	20.9	22.1	21.7	21.0
		p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001
	MPER-2/4/10	16.1	16.6	16.7	13.9	16.3	16.7
		p = 0.001	p < 0.001	p < 0.001	p = 0.025	p = 0.001	p = 0.001
	gp1ΔMPER	20.2	19.9	18.7	18.1	13.3	12.7
		p < 0.001	p = 0.001	p = 0.001	p = 0.005	p = 0.013	p = 0.02
	p17	0.0	0.0	0.0	2.8	0.0	0.0
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	p24	10.4	10.2	10.4	14.1	10.5	11.3
		p = 0.036	p = 0.031	p = 0.03	p = 0.014	p = 0.039	p = 0.036
	JR-FL gp120	9.5	9.5	9.8	7.5	8.8	9.5
		p = 0.039	p = 0.047	p = 0.04	n.s.	n.s.	p = 0.046
	BG505 trimer	5.4	5.6	5.8	0.0	5.5	6.7
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	BG505 gp140	9.3	9.8	9.8	2.8	10.1	10.7
		n.s.	p = 0.043	p = 0.04	n.s.	n.s.	p = 0.048
	RSC3	2.4	1.3	0.5	0.0	0.0	0.0
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	RSC3Δ	10.3	9.7	8.4	9.1	8.4	7.1
		p = 0.04	p = 0.049	n.s.	n.s.	n.s.	n.s.
	V3 JR-FL	8.8	9.3	9.2	6.8	8.8	7.8
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	V3 MN	6.6	7.0	6.8	3.0	6.0	5.0
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	V3 BG505	10.1	10.8	10.2	8.1	10.6	9.9
		p = 0.029	p = 0.03	p = 0.032	n.s.	p = 0.023	p = 0.04
	MPER-2/4	10.8	10.2	10.5	14.8	10.4	10.1
		p = 0.033	p = 0.039	p = 0.036	p = 0.012	p = 0.041	n.s.
	MPER-2/4/10	0.0	0.0	0.0	0.6	0.0	0.0
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	gp1ΔMPER	4.7	5.0	5.5	0.0	4.3	4.4
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
IgG3 binding	p17	12.3	11.4	10.6	7.4	10.0	9.3
		p = 0.016	p = 0.019	p = 0.029	n.s.	p = 0.044	n.s.
	p24	3.3	3.7	3.1	1.8	3.3	1.8
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	JR-FL gp120	6.0	6.1	6.1	0.0	5.4	5.5
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	BG505 trimer	9.6	9.3	8.6	4.8	8.7	8.0
		p = 0.049	p = 0.049	n.s.	n.s.	n.s.	n.s.
	BG505 gp140	13.3	12.6	11.8	4.5	12.5	11.9
		p = 0.008	p = 0.011	p = 0.017	n.s.	p = 0.014	p = 0.023
	RSC3	10.6	10.4	9.5	10.3	8.5	7.4
		p = 0.028	p = 0.038	n.s.	n.s.	n.s.	n.s.
	RSC3Δ	10.9	10.0	9.5	6.7	9.0	8.5
		p = 0.032	p = 0.047	n.s.	n.s.	n.s.	n.s.
	V3 JR-FL	15.8	15.9	15.4	12.6	14.4	13.8
		p = 0.002	p = 0.003	p = 0.004	p = 0.034	p = 0.009	p = 0.012
	V3 MN	15.7	15.7	15.3	14.6	14.8	14.3
		p = 0.001	p = 0.002	p = 0.003	p = 0.012	p = 0.004	p = 0.006
	V3 BG505	11.6	11.0	10.6	10.9	10.7	10.3
		p = 0.018	p = 0.019	p = 0.026	p = 0.031	p = 0.03	p = 0.04
	MPER-2/4	6.7	5.6	4.9	14.7	5.8	5.0
		n.s.	n.s.	n.s.	p = 0.009	n.s.	n.s.
	MPER-2/4/10	6.7	6.1	5.6	10.3	6.6	6.1
		n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
	gp1ΔMPER	9.5	8.9	9.5	5.6	7.3	6.9
		p = 0.042	n.s.	p = 0.049	n.s.	n.s.	n.s.

One-sided *P* values represent the proportion of 1,000 completely random reassignments of the *n* = 303 recipients to transmitters with maximum values greater or equal than the proportion of variability observed for the transmission pair cohort. Significant results (*P* < 0.05) are shown in red.

Experimental and computational framework for a dynamic protein atlas of human cell division

Yin Cai^{1,3,9}, M. Julius Hossain^{1,9}, Jean-Karim Hériché¹, Antonio Z. Politi^{1,4}, Nike Walther¹, Birgit Koch^{1,5}, Malte Wachsmuth^{1,6}, Bianca Nijmeijer¹, Moritz Kueblbeck¹, Marina Martinic-Kavur^{2,7}, Rene Ladurner^{2,8}, Stephanie Alexander¹, Jan-Michael Peters² & Jan Ellenberg^{1*}

Essential biological functions, such as mitosis, require tight coordination of hundreds of proteins in space and time. Localization, the timing of interactions and changes in cellular structure are all crucial to ensure the correct assembly, function and regulation of protein complexes^{1–4}. Imaging of live cells can reveal protein distributions and dynamics but experimental and theoretical challenges have prevented the collection of quantitative data, which are necessary for the formulation of a model of mitosis that comprehensively integrates information and enables the analysis of the dynamic interactions between the molecular parts of the mitotic machinery within changing cellular boundaries. Here we generate a canonical model of the morphological changes during the mitotic progression of human cells on the basis of four-dimensional image data. We use this model to integrate dynamic three-dimensional concentration data of many fluorescently knocked-in mitotic proteins, imaged by fluorescence correlation spectroscopy-calibrated microscopy⁵. The approach taken here to generate a dynamic protein atlas of human cell division is generic; it can be applied to systematically map and mine dynamic protein localization networks that drive cell division in different cell types, and can be conceptually transferred to other cellular functions.

To generate standardized, quantitative data of the dynamic 3D localization of mitotic proteins, we imaged HeLa cell lines, in which proteins were fluorescently labelled by editing the corresponding genomic locus⁶. For each protein, the cell and chromosome volumes were recorded in separate channels as spatio-temporal landmarks. We recorded mitosis in high throughput by detecting the beginning of cell division (prophase) in low-resolution images of the chromosomes, imaging the protein of interest by high-resolution 3D confocal microscopy until division was completed (Fig. 1a) and then calibrating the signal by fluorescence-correlation spectroscopy (FCS)⁷ in six nuclear and cytoplasmic positions (Fig. 1a). Calibration allowed us to convert 3D protein-fluorescence movies to time-resolved distribution maps of protein concentrations (see Methods; Fig. 1b, c). Using this automated experimental pipeline, we acquired a pilot dataset for 28 proteins, most of which were homozygously tagged with enhanced green fluorescent protein (eGFP) using zinc finger nucleases⁸ or CRISPR-Cas9 nickase-mediated⁹ genome editing, although for some genes stable integration of cDNAs or bacterial artificial chromosomes¹⁰ had to be used (see Methods; Supplementary Table 1). The time-resolved 3D distribution was recorded for 18 dividing cells per protein on average (Fig. 2a), giving us a sufficiently large dataset to develop and test our computational framework.

Although cell division is a continuous process, traditionally, mitosis is divided into five stages: prophase, prometaphase, metaphase, anaphase and telophase¹¹. Nuclear envelope breakdown and chromosome segregation mark the onset of prometaphase and anaphase, respectively, whereas the other stages are not separated by sharp kinetic boundaries.

To align the varying kinetic data from different cells (Fig. 2a), we first defined a ‘mitotic standard time’ on the basis of changes in chromosome structure. Chromosome boundaries in all imaged mitoses were automatically segmented in 3D using the landmark channel (see Methods; Fig. 2b, Extended Data Fig. 1a, b). Three geometric features were extracted from the segmented data: the distance between the two segregated chromosome masses, the total chromosome volume and the third eigenvalue of the chromosome mass (Fig. 2c). Each mitosis movie could thus be represented as a six-dimensional vector sequence of these parameters with their first derivative indicating kinetic transitions. We used a modified Barton–Sternberg algorithm with multidimensional dynamic time warping to align the vector sequences and construct a mitotic standard time reference (see Methods; Fig. 2d, Extended Data Fig. 1c, d). To discretize major transitions in chromosome structure during mitosis, we determined local maxima in the second derivative of the average feature sequences (Extended Data Fig. 2a–c). This automatically distinguished 20 mitotic stages, which we used to annotate the experimentally sampled time points of individual HeLa cells throughout the study (Extended Data Fig. 2d). We validated the generality of this approach by aligning a different human cell type (U2OS) using the same landmarks, showing the conserved nature of the mitotic transitions (Extended Data Fig. 3).

This alignment allowed us to map all cell images objectively to a standard time reference for averaging. To enable visualization, interactive navigation and analysis of all imaged protein distributions, we next computed the canonical geometry from late prometaphase (stage 7) to cytokinesis (stage 20), in which little deviation from rotational symmetry around the division axis occurs. The canonical geometry model was reconstructed from the average geometry of several hundred cells that were spatially registered for each mitotic standard stage (see Methods; Extended Data Fig. 4). Evolution of this mitotic standard geometry over the mitotic standard time defines the four-dimensional (4D) canonical mitotic cell model, enabling us to register all recorded cell divisions in space and time on the basis of their landmark channels. For each protein, we mapped each 3D stack to the corresponding mitotic standard stage (Fig. 3a) and computed 4D concentration maps representing the average behaviour of each mitotic protein. Maps of many proteins can then be freely combined (Fig. 3b)—to compare their localization patterns, dynamics and abundance—and to provide intuitive navigation of the integrated dataset, as illustrated in our web-based interactive mitotic cell atlas (http://www.mitocheck.org/mitotic_cell_atlas).

To illustrate the power of integrated data exploration for multiple proteins in the canonical model, we analysed eight mitotic chromosome structure proteins (Extended Data Fig. 5a, b). Plotting the total number of proteins found on mitotic chromosomes and in the daughter nuclei against the mitotic standard time allowed for quantitative comparison of protein dynamics (see Methods; Fig. 3c, d), which revealed that the amount of most chromosomal proteins bound to chromatin

¹European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. ²Research Institute of Molecular Pathology (IMP), Vienna, Austria. ³Present address: Roche Diagnostics, Waiblingen, Germany. ⁴Present address: Max Planck Institute for Biophysical Chemistry, Goettingen, Germany. ⁵Present address: Max Planck Institute for Medical Research, Heidelberg, Germany. ⁶Present address: Luxendo GmbH, Heidelberg, Germany. ⁷Present address: Genos, Glycoscience Research Laboratory, Zagreb, Croatia. ⁸Present address: Stanford School of Medicine, Stanford, CA, USA. ⁹These authors contributed equally: Yin Cai, M. Julius Hossain. *e-mail: jan.ellenberg@embl.de

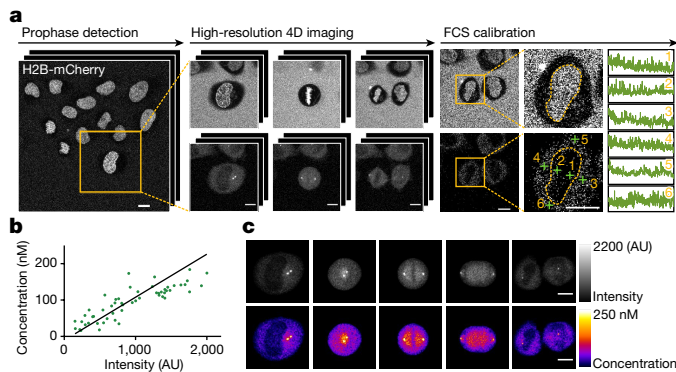


Fig. 1 | Quantitative imaging of mitotic proteins. **a**, Automatic calibrated 3D live confocal imaging pipeline. Cells in prophase were identified by online classification, imaged through mitosis in the channels of the landmarks and proteins of interest, and measured by FCS at selected positions. **b**, The local protein concentrations determined by FCS fitting linearly correlate with the background subtracted image intensities at the corresponding positions (data acquired on the same day and using the same microscope are shown). AU, arbitrary units. **c**, Example cell showing concentration maps resulting from FCS-based intensity calibration (mean *z*-projection). Scale bars, 10 μ m. Data shown in **a–c** are for H2B-mCherry mNEDD1-LAP (eGFP) and are representative of 92 independent experiments performed with 28 different cell lines.

in metaphase is within the same order of magnitude (3,000–26,000 per chromosomal volume), except for TOP2A which shows an abundance 25 times higher, potentially suggesting a structural rather than a purely enzymatic role. We found that cohesins slowly and progressively dissociate from chromatin in early mitosis (RAD21, STAG1 and STAG2), with a final more-rapid release of the approximately 3,000 remaining cohesin complexes before the onset of anaphase. This observation indicates that no more than 100 cohesins—bound mostly at the centromere (see Methods and previously published work¹²)—are sufficient to connect the sister chromatids on an average human chromosome. Notably, the cohesins bound to mitotic chromosomes consisted of equal amounts of two isoforms containing the HEAT repeat subunits STAG1 or STAG2 (Fig. 3d), contrasting with the situation in interphase nuclei in which STAG2-containing complexes dominate¹³. Furthermore, we observed that a significant amount of STAG2 ($P < 0.025$, paired Wilcoxon signed-rank test), but not of the kleisin subunit RAD21, rebound chromosomes in anaphase, suggesting a potential non-cohesive function of STAG2 during mitotic exit (Fig. 3c, d, Extended Data Fig. 5b). In contrast to the complete dissociation of most cohesins, about 17,000 molecules of the chromatin organizer CTCF remained associated with the genome throughout mitosis¹⁴, consistent with a ‘bookkeeping’ mechanism of interphase chromatin architecture. Once chromosome segregation was initiated, KIF4A, TOP2A and CTCF further accumulated on chromatin in anaphase, suggesting a role in the maximal shortening of chromosome arms in anaphase¹⁵. During nuclear reformation, the cytoplasmic pool of mitotic chromosome proteins showed an ordered entry, as well as decreasing rates of import. CTCF was reimported first with the highest rate (391 proteins per second), and then the cohesin subunits STAG1 and RAD21 were simultaneously imported (352 and 239 proteins per second, respectively), whereas STAG2 and WAPL enter the nucleus later and at a lower rate (69 and 89 proteins per second, respectively). This finding confirms that mitotic decondensation proceeds in the presence of CTCF and—subsequently—STAG1-containing cohesin complexes, but before WAPL and STAG2 are present⁵². In addition to the chromosomal proteins, we also explored assembly of the nuclear pore complex (NPC) during late anaphase (Extended Data Fig. 5c). Consistent with previous observations^{16,17}, we found that cytoplasmic ring components (NUP107 and NUP214) assembled early, but also found that nuclear basket and cytoplasmic filament proteins (TPR and RANBP2) assembled much later, at a time when the import of CTCF was already completed. This suggests that nuclear and cytoplasmic

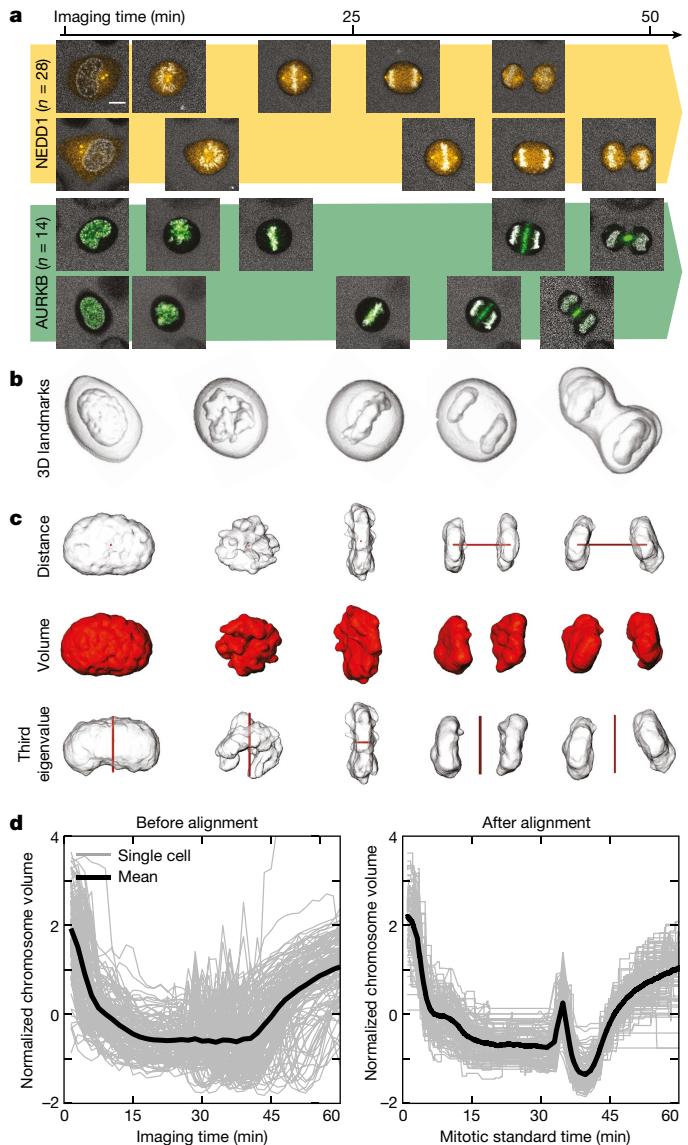


Fig. 2 | Modelling of mitotic standard time. **a**, Individual cells have different mitotic spatio-temporal dynamics. Scale bar, 10 μ m. **b**, Cellular and chromosomal volumes were segmented from the landmark channels. **c**, Three morphological features (in red) were extracted from the chromosomal volume. **d**, Mitotic standard time was generated in the feature space by multiple sequence alignment visualized here in the feature dimension describing chromosomal volume. The alignment of 132 cells from 20 independent experiments is shown.

filaments of the NPC are not required for the rapid import of nuclear proteins, which is needed for the re-establishment of interphase genome architecture (Extended Data Fig. 5d).

To comprehensively investigate which proteins work together, and when and where inside the cell they interact, we transformed their spatial distribution into numerical features. We used a segmentation-free approach, formulated on the basis of a speeded-up robust features (SURF) detector¹⁸, to extract so-called interest point clusters (see Methods; Fig. 4a, Extended Data Fig. 6a, b) transforming each 3D movie into a sequence of 100-dimensional feature vectors. By averaging the feature vectors of all images of the same protein and mitotic standard stage, the dynamic distribution of all proteins in our dataset could then be represented as a third-order tensor of size $28 \times 100 \times 20$ (proteins \times features \times stages). Soft clustering with non-negative tensor factorization detected seven clusters of dynamic protein localization patterns (see Methods, Fig. 4a, Extended Data Fig. 6c). The identity of the proteins in each pattern revealed a correspondence between these

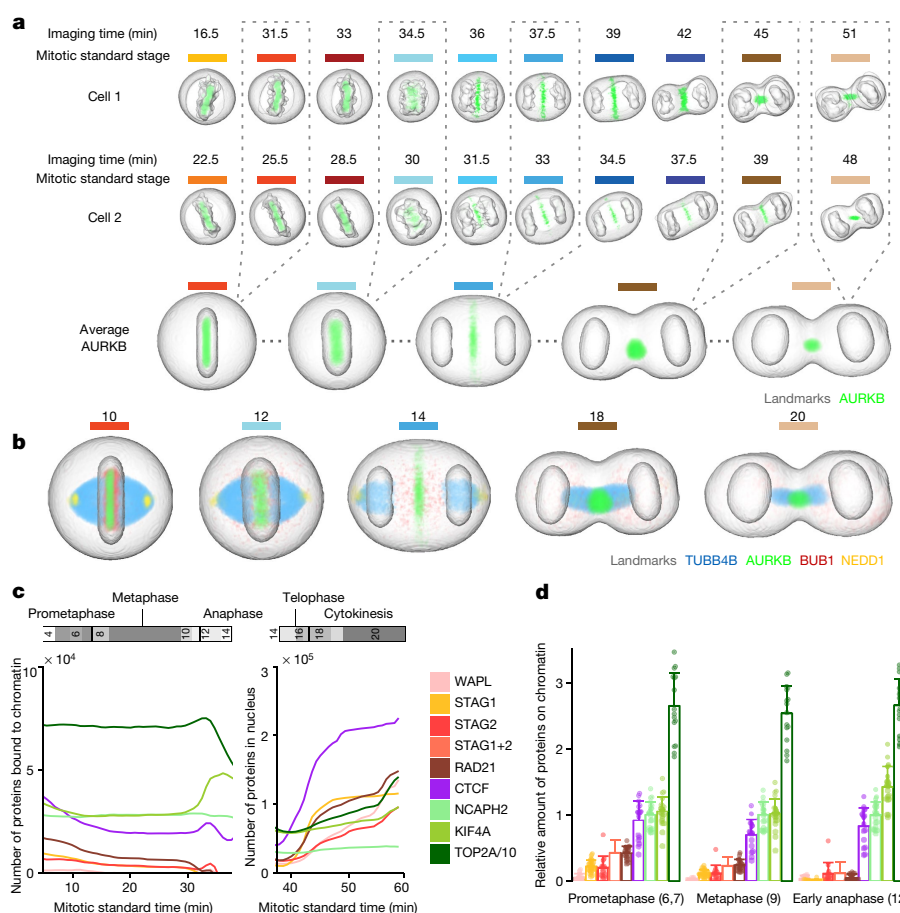


Fig. 3 | Visualization of 4D protein-distribution maps. a, Through averaging of a large number of cells, models were generated for all mitotic standard stages with symmetrical geometries. Example image sequences were registered to the standard space of the corresponding mitotic standard stage. A distribution map over time was then generated for each protein by averaging through multiple cells. Coloured lines indicate mitotic stages. **b**, Average distributions of four proteins are displayed in different mitotic stages. **c**, Amount of chromatin-bound and nuclear

molecules for eight chromatin organizers. **d**, Fraction of chromatin-bound proteins relative to NCAPH2. Shown are the single cell values (dots) and the mean + s.d. The sum of STAG1 and STAG2 (STAG1 + 2) was calculated from the mean + s.d. of STAG1 and STAG2 data. In **c** and **d**, TOP2A has been scaled down by a factor of ten for visualization. Note, reported numbers represent monomers; dimers (for example, TOP2A) would result in a 50% reduced abundance of complexes.

statistically defined clusters and major mitotic organelles and structures (for example, CENPA identifies centromeres and RACGAP1 identifies the midplane in late mitotic stages; Extended Data Fig. 7), validating our unsupervised approach. As our clustering assigns the fraction of a mitotic protein to each pattern over time, it reliably deals with promiscuous proteins present in multiple sub-cellular structures. Linking proteins with similar patterns at each time point allowed us to derive a dynamic multigraph that showed the dynamic network of protein colocalization and highlighted the activities of different compartments over time, allowing us to predict where and when proteins were most likely to interact (Fig. 4b). Results from the above clustering can be used to generate hypotheses that can be visualized in the canonical cell model. For instance, the temporal evolution of the mitotic kinase AURKB and its regulator CDCA8 (also known as borealin), predicts that the two proteins relocate to the midplane (Fig. 4a, purple in right panel) in different proportions and with different kinetics. As the two proteins are known to be present in a 1:1 ratio in the chromosome passenger complex¹⁹, this observation suggests that a fraction of AURKB in the midplane is not part of the complex. Exploring the 4D localization of CDCA8 and AURKB in the mitotic cell atlas (Extended Data Fig. 8a, b) revealed that although these two proteins partially colocalize at the midbody, AURKB exhibits an additional localization in an equatorial cortical ring that contracts as mitosis progresses. This novel localization of the mitotic kinase AURKB suggests that it is an integral part of the contractile cytokinetic ring. Although unexpected, this observation is

consistent with the function of AURKB in cytokinesis^{20–22}. This raises the very interesting possibility that the midplane and cytokinetic ring pools of AURKB have different functions for central spindle architecture and cytokinesis, respectively, during mitotic exit.

As our clustering of dynamic localization patterns does not yield pure sub-cellular compartments as defined ultrastructurally or by fractionation a priori, we developed a supervised machine-learning approach to define subcellular structures on the basis of known resident proteins of six compartments and organelles that are relevant for mitosis: chromosomes, nuclear envelope, kinetochores, spindle, centrosomes and midbody (see Methods, ‘Analysis of protein localization kinetics using supervised annotation’). Using the interest point cluster features as input, we trained a multivariate linear regression model that could assign the amount of a protein of interest present in each of the six reference compartments (Extended Data Fig. 8c, d). This allowed us to quantitatively compare the subcellular fluxes between these compartments for all proteins (Extended Data Fig. 9).

Mining this dataset is a powerful approach for dissecting dynamic multimolecular events inside living cells, such as the assembly or disassembly of organelles. As an example, we calculated the number of all imaged proteins localized to kinetochores to investigate the disassembly of this large supramolecular complex, a process that is essential for cell division. The data allowed us to determine that kinetochore disassembly starts in early metaphase with the dissociation of BUB1B and PLK1 and of BUB1, AURKB, MIS12 and CDCA8 in late metaphase

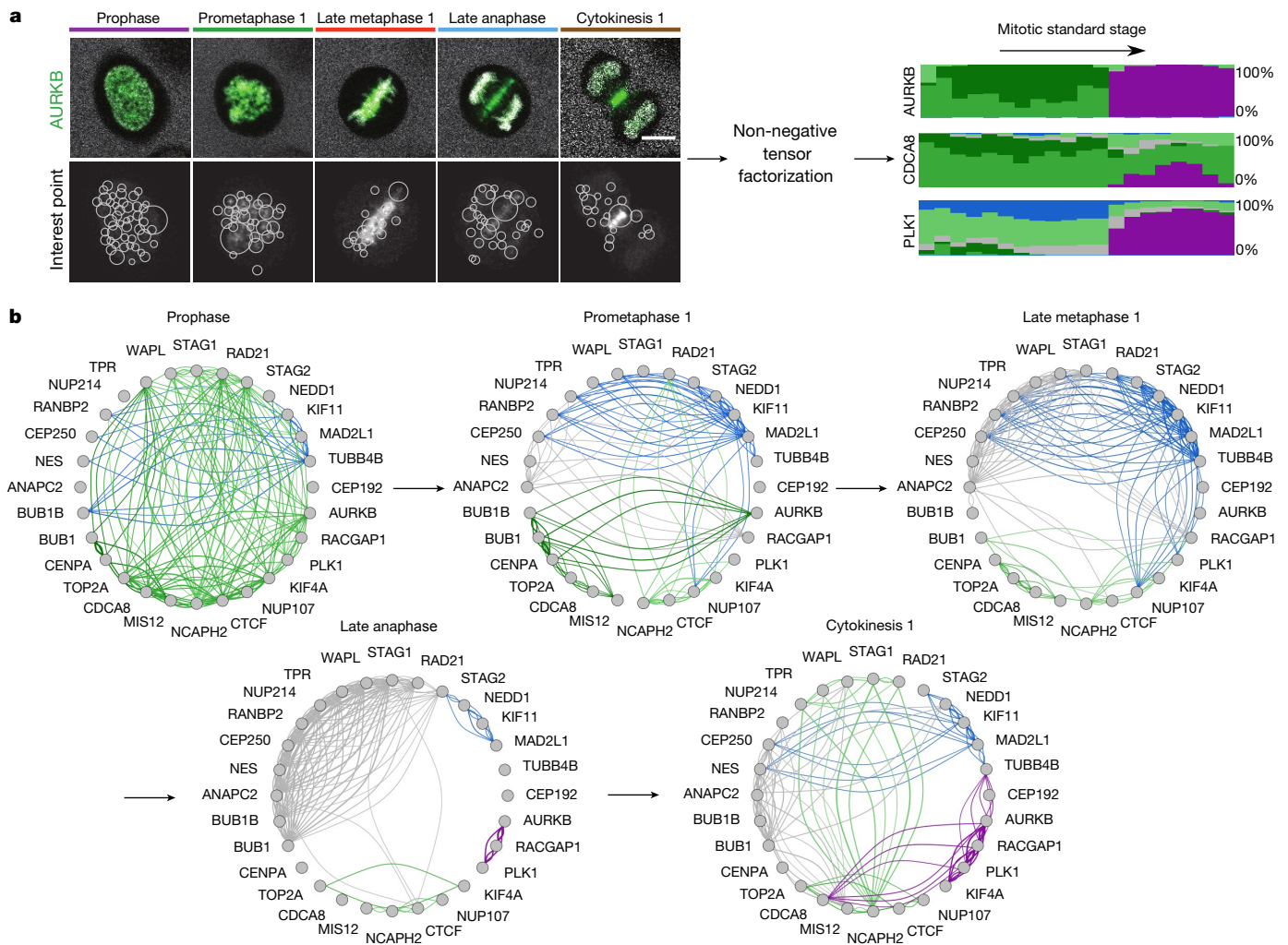


Fig. 4 | Identification of dynamic protein clusters. a, SURF interest points were detected and assigned to one of 100 clusters of similar interest points (bottom left, circled). Non-negative factorization of the data tensor of 28 proteins \times features \times mitotic stages produced a non-negative tensor of reduced dimension, the entries of which can be interpreted as the fraction of protein belonging to each cluster over time (right panel, each cluster is represented by a different colour and the height of a coloured

bar at a given mitotic stage represents the fraction of the protein in the corresponding cluster at this stage). Scale bar, 10 μ m. **b**, Dynamic multigraph of protein colocalization, shown for five stages. Each edge colour corresponds to a localization cluster (as in **a**) and the edge thickness corresponds to the product of the fraction of linked genes in the corresponding cluster and can be loosely interpreted as a probability of interactions.

(Extended Data Fig. 8d). In addition, this analysis showed that the stoichiometry of these proteins before disassembly differs up to tenfold, and that their maximal dissociation rates span over an order of magnitude (ranging from 10 to 131 molecules per second). The predicted number of approximately 300 CENPA molecules per centromere was consistent with data from biochemical methods²³ and the predicted disassembly order concurs with reports of the late dissociation of MIS12²⁴ (Extended Data Fig. 8e).

Our automatic assignment of protein quantities to cellular organelles allowed us to determine the exact timing, stoichiometry and dissociation rates for multiple mitotic proteins that reside dynamically on kinetochores.

With this study we provide an integrated experimental and computational framework to build a comprehensive and quantitative 4D model of the mitotic protein localization network in a dividing human cell. Our model provides a standardized, yet dynamic, spatio-temporal reference system for the mitotic cell that can be used to integrate quantitative information on any number of protein distributions sampled in thousands of different single-cell experiments. Using a pilot dataset, we illustrate the power of this model by mining the data to automatically define dynamic localization patterns to subcellular structures as well as predicting the order, stoichiometry and rates of assembly and

disassembly of sub-cellular organelles. This quantitative information on protein localization in living cells provides greater insights into protein dynamics and interactions at relevant temporal resolutions and supports simulations of mitotic processes. Our computational model underpins an interactive 4D atlas of the human mitotic cell, which allows the visualization of multiple protein dynamics with a spatial and temporal resolution and continuity that are currently very difficult or impossible to reach with multi-colour live-cell imaging over the duration of mitosis. We illustrate with mitotic chromosome formation, kinetochore disassembly, NPC assembly and cytokinesis how the knowledge gained through the exploration and mining of the atlas data can be used to formulate new mechanistic hypotheses about the function of proteins inside the cell. The concept of standardizing the spatio-temporal cellular context for analysing dynamic protein distributions to understand cellular processes as presented here is generic and we envision its adaptation to other essential biological functions such as cell migration or cell differentiation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0518-z>.

Received: 7 March 2016; Accepted: 25 July 2018;
Published online: 10 September 2018

1. Rodriguez-Bravo, V. et al. Nuclear pores protect genome integrity by assembling a premitotic and Mad1-dependent anaphase inhibitor. *Cell* **156**, 1017–1031 (2014).
2. Dick, A. E. & Gerlich, D. W. Kinetic framework of spindle assembly checkpoint signalling. *Nat. Cell Biol.* **15**, 1370–1377 (2013).
3. Ghongane, P., Kapanidou, M., Asghar, A., Elowe, S. & Bolanos-Garcia, V. M. The dynamic protein Knl1—a kinetochore rendezvous. *J. Cell Sci.* **127**, 3415–3423 (2014).
4. Fink, J. et al. External forces control mitotic spindle positioning. *Nat. Cell Biol.* **13**, 771–778 (2011).
5. Politi, A. Z. et al. Quantitative mapping of fluorescently tagged cellular proteins using FCS-calibrated four-dimensional imaging. *Nat. Protocols* **13**, 1445–1464 (2018).
6. Koch, B. et al. Generation and validation of homozygous fluorescent knock-in cells using CRISPR-Cas9 genome editing. *Nat. Protocols* **13**, 1465–1487 (2018).
7. Magde, D., Elson, E. & Webb, W. W. Thermodynamic fluctuations in a reacting system measurement by fluorescence correlation spectroscopy. *Phys. Rev. Lett.* **29**, 705–708 (1972).
8. Hockemeyer, D. et al. Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nat. Biotechnol.* **27**, 851–857 (2009).
9. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
10. Poser, I. et al. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods* **5**, 409–415 (2008).
11. Alberts, B. et al. *Molecular Biology of the Cell* 4th Edition (Garland Science, New York, 2002).
12. Landry, J. J. M. et al. The genomic and transcriptomic landscape of a HeLa cell line. *G3 (Bethesda)* **3**, 1213–1224 (2013).
13. Holzmann, J., Fuchs, J., Pichler, P., Peters, J. M. & Mechtler, K. Lesson from the stoichiometry determination of the cohesin complex: a short protease mediated elution increases the recovery from cross-linked antibody-conjugated beads. *J. Proteome Res.* **10**, 780–789 (2011).
14. Burke, L. J. et al. CTCF binding and higher order chromatin structure of the H19 locus are maintained in mitotic chromatin. *EMBO J.* **24**, 3291–3300 (2005).
15. Mora-Bermúdez, F., Gerlich, D. & Ellenberg, J. Maximal chromosome compaction occurs by axial shortening in anaphase and depends on Aurora kinase. *Nat. Cell Biol.* **9**, 822–831 (2007).
16. Otsuka, S. et al. Nuclear pore assembly proceeds by an inside-out extrusion of the nuclear envelope. *eLife* **5**, e19071 (2016).
17. Otsuka, S. et al. Postmitotic nuclear pore assembly proceeds by radial dilation of small membrane openings. *Nat. Struct. Mol. Biol.* **25**, 21–28 (2018).
18. Bay, H., Ess, A., Tuytelaars, T. & Van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **110**, 346–359 (2008).
19. Carmenta, M., Wheelock, M., Funabiki, H. & Earnshaw, W. C. The chromosomal passenger complex (CPC): from easy rider to the godfather of mitosis. *Nat. Rev. Mol. Cell Biol.* **13**, 789–803 (2012).
20. Hauf, S. et al. The small molecule Hesperadin reveals a role for Aurora B in correcting kinetochore-microtubule attachment and in maintaining the spindle assembly checkpoint. *J. Cell Biol.* **161**, 281–294 (2003).
21. Steigemann, P. et al. Aurora B-mediated abscission checkpoint protects against tetraploidization. *Cell* **136**, 473–484 (2009).
22. Isokane, M. et al. ARHGEF17 is an essential spindle assembly checkpoint factor that targets Mps1 to kinetochores. *J. Cell Biol.* **212**, 647–659 (2016).
23. Bodor, D. L. et al. The quantitative architecture of centromeric chromatin. *eLife* **3**, e02137 (2014).
24. Gascoigne, K. E. & Cheeseman, I. M. CDK-dependent phosphorylation and nuclear exclusion coordinately control kinetochore assembly state. *J. Cell Biol.* **201**, 23–32 (2013).
52. Zuin, J. et al. A cohesin-independent role for NIPBL at promoters provides insights in CdLS. *PLoS Genetics* **10**, e1004153 (2014).

Acknowledgements We thank T. Hyman and I. Poser for donating multiple mouse BAC protein–GFP cell lines and T. Hirota for giving us the eGFP–CENPA cell line. The automatic imaging would not have been possible without R. Höfler and D. W. Gerlich, who developed the Micronaut software. We thank all members of the Ellenberg and Peters laboratories for support, especially M. Isokane, M. J. Roberti, J. Mergenthaler, S. Otsuka, W. Tang and D. Cisneros for generating cell lines, reagents and constructs. We thank A. Callegari for supporting the U2OS data generation. We also thank W. Huber, B. Fischer, B. Klaus and L. P. Coelho for discussions, the EMBL mechanical and electronic workshop, the EMBL advanced light microscopy facility, the EMBL flow cytometry core facility and the IMP BioOptics Facility for their support. This study has benefited from the collaboration with Carl ZEISS Jena, especially with T. Ohrt. The work was supported by grants from EU-FP7-MitoSys (Grant Agreement 241548) to J.E. and J.M.P., EU-FP7-SystemsMicroscopy NoE (Grant Agreement 258068), EU-H2020-iNEXT (Grant Agreement 653706) and the 4D Nucleome/4DN National Institutes of Health common fund (5 U01 EB021223-04 / 8 U01 DA047728-04) all to J.E., as well as by the European Molecular Biology Laboratory (Y.C., M.J.H., J.-K.H., A.Z.P., N.W., B.K., M.W., B.N., M.K., S.A. and J.E.). Y.C. and N.W. were also supported by the EMBL International PhD Programme (EIPP). Research in the laboratory of J.M.P. was further supported by Boehringer Ingelheim, the Austrian Science Fund (FWF special research program SFB F34 ‘Chromosome Dynamics’ and Wittgenstein award Z196-B20), the Austrian Research Promotion Agency (Headquarter grants FFG-834223 and FFG-852936) and the European Research Council (ERC) under the European Union Horizon 2020 research and innovation programme (Grant Agreement 693949).

Reviewer information Nature thanks R. Murphy, J. Swedlow and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.K., B.N., M.K., R.L., M.M.K. and N.W. constructed and validated GFP knock-in cell lines. Y.C., A.Z.P. and M.W. developed the imaging pipeline. Y.C., A.Z.P. and N.W. performed the calibrated live-cell imaging. Y.C., M.J.H., J.-K.H. and A.Z.P. developed the computational analysis pipeline. Y.C., J.-K.H., M.J.H., A.Z.P., S.A. and J.E. wrote the manuscript. J.M.P. coordinated the MitoSys consortium and supervised part of the work. J.E. supervised the work overall and originally conceived the project.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0518-z>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0518-z>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to J.E.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture. HeLa Kyoto cells (RRID: CVCL_1922) were a gift from S. Narumiya. These cells were authenticated by whole-genome sequencing. HeLa Kyoto cells were cultured in high glucose Dulbecco's modified Eagle's medium (DMEM; Life Technologies) supplemented with 10% (v/v) fetal bovine serum (FBS), 100 units/ml penicillin, 0.1 mg/ml streptomycin, 2 mM glutamine and 1 mM (v/v) sodium pyruvate at 37°C and 5% CO₂. Depending on the genetic modification, one or more of the following antibiotics were supplied to the culture at the stated final concentration: geneticin (Life Technologies) 500 µg/ml, hygromycin B (Invitrogen) 200 µg/ml or puromycin (Invitrogen or Calbiochem) 0.5 µg/ml. Once the cells reached 80–90% confluence, they were passaged and only a fraction of the cells were cultured in a fresh dish. U2OS cells were obtained from the ATCC (HTB-96) and were not further authenticated. The U2OS cells were cultured in McCoy's 5A medium (Sigma-Aldrich) supplemented with 10% (v/v) FBS, 100 units/ml penicillin, 0.1 mg/ml streptomycin, 2 mM glutamine, 1 mM (v/v) sodium pyruvate, and 1% (v/v) MEM non-essential amino acids (Gibco). All cells tested negative for mycoplasma contamination.

Cell modification. HeLa Kyoto cells were used for genetic modifications and imaging. HeLa Kyoto cells are hypotriploid with on average 64 chromosomes, thus during mitosis the cells have on average $64 \times 2 = 128$ kinetochores¹². The cell lines that have been generated for this project or previously^{16,25–31} are listed in Supplementary Table 1 with their providers indicated. Several cell lines were generated within this project as follows: the cell lines expressing PLK1–mGFP, CEP192–mGFP and mGFP–NUP107 were generated using the Zinc finger nuclease pipeline as previously described²⁸. Zinc finger nucleases were purchased from Sigma-Aldrich and DNA-binding sequences are listed in Supplementary Table 2. The other genome-edited cell lines were generated using the CRISPR–Cas9 system⁹ on the basis of the paired Cas9 nickase approach. For these cell lines, both gRNAs (Supplementary Table 2) and the donor plasmid were designed on the basis of ENSEMBL release 75 and transfected together into HeLa Kyoto cells with jetPrime (Polyplus), FuGENE (Promega) or Lipofectamine2000 (Thermo Fisher) according to the manufacturers' instruction. A single clone was selected using our previously developed validation pipeline^{6,28}. For 4 out of 20 genome-edited cell lines (BUB1B–eGFP, TPR–mGFP, mGFP–NUP107 and CEP192–mGFP), we detected a band of the size of free GFP in western blots (anti-GFP, Roche 11814460001). Therefore, for these cell lines, the total amount of freely diffusing tagged protein may be overestimated. To label the chromosomal volume, an H2B–mCherry³² cDNA was transfected into some genome-edited cell lines with Eugene6 (Promega) according to the manufacturer's instructions. The pmeGFP2–N1–NES construct was generated by sub-cloning two tandem repeats of monomer eGFP (mEGFP2)³³ and the nuclear export signal (NES) of MAPKK (NLVDLQKKLEELDEQQ) into the pEGFP–N1 vector (Clontech Laboratories). The pmeGFP2–N1–NES construct was transfected into HeLa Kyoto cells and cells with stable expression were selected by culturing with the appropriate antibiotics. Single cells or a cell population with the desired expression level were collected for imaging by fluorescence-activated cell sorting (FACS, performed by the EMBL Heidelberg facility).

Calibrated fluorescence confocal microscopy. Confocal microscopy was performed on Zeiss LSM780, Confocor 3, laser scanning microscopes using 40×, NA 1.2 water DIC Plan-Apochromat objectives and the Gallium arsenide phosphide (GaAsP) detectors equipped with an incubation chamber (EMBL workshop). Cells were imaged at 37°C in a CO₂-independent medium (Life Technologies) fluorescently coloured with 500 kDa Dextran (Life Technologies)–DY481XL (Dyomics) produced in house. Time-lapse imaging was performed using the ZEN 2012 software as well as in-house software applications (see previously published work³ for a software description). The acquisition was supported by an in-house-developed objective cap and a water pump, such that water drops were regularly supplied to the objective–sample interface. Before starting imaging, a number of positions were selected manually. During live-cell imaging, the microscope determined the focus automatically by performing line-scan imaging of the reflection signal of the 633-nm laser. The vertical position of the glass bottom was determined as the position with the maximum reflection intensity, and used as reference for acquiring a volume of the specimen at a particular depth.

The imaging workflow was set-up using the VBA Zeiss Macro MyPic (<https://git.embl.de/grp-ellenberg/mypic>). HeLa Kyoto cell lines with H2B–mCherry were imaged live using an excitation laser at 561 nm every 5 min for about 16 h on the Zeiss 780 microscopy system. Three confocal planes were acquired at a resolution of 0.32 µm in *x* and *y* and 2.5 µm in *z*. Images were projected in *z* by taking the maximum intensity value. Images of the H2B signal were analysed while in progress by the Micronaut software (Gerlich laboratory, IMBA), using a support vector machine classifier that was trained beforehand (with the software CellCognition, <https://www.cellcognition.org/>) to distinguish between cells

in interphase, prophase, mitosis (prometaphase–telophase) and artefacts (apoptosis, on the border of the imaging field, out of focus, too-low expression). The classification score for prophase, interpreted as the probability of a cell being in the class of interest, was output, and a pre-defined threshold was used to make a decision on whether imaging setups for mitotic cell acquisition should be activated. Depending on how different the H2B–mCherry expression levels of the sample were from the training set, the threshold on class probability was set between 0.85 and 0.96. Once a prophase cell was found, it was then imaged using a different imaging setup. For our purpose, mitotic cells were imaged live every 90 s for 31 *z*-planes with a spatial resolution of 0.25 µm in *x–y* and 0.75 µm in *z* with a 488-nm laser (high expression of H2B–mCherry allowed it to be excited at 488 nm and produce adequate signal). For cells not expressing H2B–mCherry, we used SiR-DNA to stain the chromatin (Spirochrome, final concentration 50 nM added 2 h before imaging). Cells were imaged with the 633-nm laser (3 confocal planes every 7.5 min at the same resolution as for H2B–mCherry) and processed as for H2B–mCherry. For U2OS cells, the chromatin was stained with 200 nM SiR-DNA. To increase the incorporation of SiR-DNA, the imaging medium of U2OS contained 1 µM verapamil (Spirochrome).

The signal from the GaAsP detector was separated into three channels: GFP, varied from 490 to 552 nm, depending on the expression level, to avoid detection saturation; mCherry, 587–621 nm (Extended Data Fig. 1a, top row), and Dy-481XL, 622–695 nm (Extended Data Fig. 1a, second row). For SiR-DNA we used 622–695 nm. Once the mitosis was recorded for 40 frames, a single-plane image was then acquired at 2.5 µm above the cover glass surface. Using an adaptive feedback microscopy Fiji macro (https://git.embl.de/grp-ellenberg/adaptive_feedback_mic_fiji), the image was thresholded using a previously developed method³⁴ and the object closest to the image centre with a proper size was selected as one of the two daughter nuclei in the cell of interest. The segmented nuclear boundary was fitted with an ellipse. FCS measurements were performed with the 488-nm laser and the avalanche photodiode (APD) detector at 505–540 nm at two positions within, and four around, the nucleus with a distance of 2 µm to the ellipse boundary for 30 s each. A manual quality control was then performed. Videos of cells with no expression of the protein of interest, with wrongly selected FCS positions (for example, outside of the cell) or without anaphase onset were excluded from further processing. A total of 499 cells were retained with an average of 18 cells per protein (ranging from 10 to 35 cells per protein).

Segmentation of the landmarks. A fully automated computational pipeline was implemented in MATLAB (MATLAB R2017a, MathWorks) to segment and track cells of interest and reconstruct chromosomal and cellular surfaces (https://git.embl.de/grp-ellenberg/mitotic_cell_atlas). The pipeline was composed of three major steps: segmentation of chromosomal volume, segmentation of cell volume and extraction of parameters from the landmarks geometry. Chromosomal regions were segmented from the mCherry channel, which had high H2B–mCherry signal and very low Dextran–Dy481XL intensity (Extended Data Fig. 1a, top row) or from the SiR-DNA channel, which had no crosstalk from other channels. To perform isotropic 3D image processing, adjacent *x–y* planes were linearly interpolated along the *z* direction. A 3D Gaussian filter was applied to reduce the effects of noise. To detect chromosomal regions, the filtered image stack was binarized first using a multi-level thresholding method as previously described³⁵. In this approach, a global Otsu threshold³⁶ was determined for the entire stack and the threshold was then adapted for each 2D slice, validated by the connectivity of binary components in 3D. Tiny connected components were removed from the binary image leaving only chromosomal components from all cells in the imaging field. All components were used as seeds for the detection of the cell boundary in a later stage. The connected chromosomal volume in the *x–y* centre of the first frame identified the cell of interest owing to the centring step in the imaging pipeline. The cell of interest was tracked sequentially through the entire image sequence using a nearest-neighbour approach. At each time point, an event of chromosome segregation was also probed by analysing the chromosomal volume around the tracked location. Once segregation was detected, both daughter nuclei were tracked in the subsequent frames (Extended Data Fig. 1a, third row).

The cell region was segmented from the Dy481XL channel showing the cell-free regions in high intensities and histone signal with low intensities (Extended Data Fig. 1a, second row). Upon interpolation and filtering as for the chromosomal segmentation, a ratio image was created by dividing the filtered image stack of the mCherry channel by that of the Dy481XL channel to diminish bleed-through signal from the H2B–mCherry channel. The ratio image stack was then binarized as described above. When using SiR-DNA, there was no bleed-through in the Dy481XL channel when excited at 488 nm. In this case, the Dy481XL was directly binarized. To separate individual cell regions, the previously detected nuclear seeds were used, taking into account the fact that each cell region can have only one or two chromosomal volumes. This was implemented by applying a marker-controlled watershed algorithm³⁷. To obtain a better separation between touching cells, the algorithm was applied on the distance-transformed image that made

use of the geometric properties of the cell surface. The cell region of interest was defined by taking the connected region(s) containing the detected chromosomal volume(s) of interest (Extended Data Fig. 1a, bottom row).

The chromosomal mass at each time point was represented by its three orthogonal eigenvectors and associated eigenvalues, where the eigenvector with the largest eigenvalue represented the longest elongated axis of the chromosomal volume. Metaphase frames were automatically detected on the basis of the low value of the smallest eigenvalue of the chromosomal volume. The division axis for metaphase cells was then predicted by taking the eigenvector having the minimum eigenvalue. By definition, this vector is always orthogonal to the metaphase plate. Using the predicted axis in the first and last metaphase frame, axes for the remaining frames were propagated backwards and forwards for stages before and after metaphase, respectively, where the eigenvector with the smallest discrepancy in angle to the axis predicted for the adjacent frame was used (Extended Data Fig. 1b). For further analysis, the plane orthogonal to the division axis going through the centroid of both daughter nuclei was predicted as the midplane. Segmented landmarks were 3D reconstructed and visualized for quality control. Cells with few time points that were over- or under-segmented were reprocessed with different parameters or by using the results of correctly segmented adjacent time points as constraint.

Image processing and calibration. Image processing and calibration were performed according to previously published methods⁵. Before each calibrated live-cell confocal microscopy experiment, the focal volume was calibrated using a 10–50-nM solution of Alexa488 (Life Technologies), and single mGFP brightness was calibrated by performing FCS measurements on HeLa Kyoto cells expressing mGFP²⁸. All FCS measurements were processed using Fluctuation Analyzer³⁸. Autocorrelation functions of dye solutions were fitted using a one-component diffusion model with triplet-like blinking, and measurements of fluorophore-fused proteins were fitted using a two-component anomalous diffusion model with fluorescent protein-like blinking. The effective confocal volume was calculated from

$$V_{\text{eff}} = (4 \cdot \pi \cdot D \cdot \tau)^{\frac{3}{2}} \cdot \kappa$$

in which D was the diffusion coefficient of the Alexa488, which is $464 \mu\text{m}^2/\text{s}$ at 37°C , and κ the structural parameter. The averaged time passing through the confocal volume τ and the structural parameter (typically between 4–7) were fitted to the auto correlation function of Alexa488. The number of fluorescent molecules within a confocal volume was calculated by multiplying the fitted number of molecules (N) with correction factors for background and photobleaching³⁸. As proteins might exist in complexes with multiple molecules, a count per molecule (CPM) value was used to correct the number of molecules. As reference, the CPM value of mGFP was used as measured in the HeLa Kyoto cells expressing mGFP, in which the mean value of all mGFP measurements was taken. If the CPM of a measurement of a fusion protein of interest within a cell was larger than that of the mGFP, the fitted number of molecules was corrected by multiplication with the ratio between the two. Finally, the local concentration of the measured protein was determined as the corrected number of molecules divided by the effective confocal volume. As a quality control of the FCS measurements, we pre-defined thresholds and deleted data points with too low coefficient of variation R^2 or APD counts or too high fitting χ^2 or bleaching or outlier CPM values.

The calibration of the image acquired with the GaAsP detector was based on the assumption of linear correlation between the local protein concentration and the eGFP intensity, which we could verify (Fig. 1b). The averaged intensity of the GFP channel in cell-free areas was considered as background. For all measurement points the coefficient ρ between local protein concentration and background-corrected imaging intensity, mean filtered with a 9×9 -pixel window to avoid noise, was calculated by performing a linear regression. The 3D protein concentration map was generated by multiplying the pixel intensities with the linear coefficient ρ . The protein number in each voxel was obtained by multiplying the concentration with the voxel volume. The absolute protein abundance could be calculated by summing up the map over the cell volume. After nuclear envelope breakdown (NEBD) and before the nuclear envelope reforms, we estimated the number of proteins bound to chromatin by subtracting the cytoplasmic average concentration (representing the background concentration of proteins that freely diffuse between the cytoplasmic and chromosome volume) from the average protein concentration on the chromosome mask. Finally, to obtain the number of proteins, the concentration difference was multiplied by the number of voxels of the chromosome mask and the voxel volume.

To assess the accuracy of our quantitative measurements, we compared our data for nucleoporins (NUP107, NUP214, TPR and RANBP2) to expected numbers calculated from the known number of nuclear pores complexes (NPC) per cell and known protein stoichiometry in each NPC. The HeLa Kyoto cell line used in this study has about 10,000 NPCs in interphase before NEBD¹⁶. Assuming a nucleoporin (NUP) stoichiometry as reported³⁹ (32 NUPs/NPC for NUP107, TPR and RANBP2; 16 NUPs/NPC for NUP214), we can compute the number of NUPs

present on the nuclear envelope. Considering a free pool of nucleoporins in the cytoplasm that is included in our measurements, the ratio of our measurements over expected numbers on the nuclear envelope should be greater than 1. We find that this ratio is on average 1.2 for all 4 NUPs, underlining the consistency of our measurements with established protein numbers by orthogonal methods.

Modelling of the mitotic standard time. The mitotic standard time was modelled in a six-dimensional feature space using three morphological features of the chromosomal volume: the distance between the two daughter nuclei, the total volume and the third eigenvalue (Fig. 2c) and their first derivatives. The model was generated by aligning 132 mitotic image sequences using the Barton-Sternberg multiple sequence alignment algorithm⁴⁰ (Extended Data Fig. 1d). The two sequences with the smallest distance to the average of all sequences were selected to initiate the alignment and each of the remaining sequences was then aligned to the average among all aligned sequences. The alignment was implemented as a modified multi-dimensional dynamic time warping⁴¹, in which the total Euclidean distance over time between the pair of sequences was used as the objective of the optimization. The timeline of the averaged sequences was calculated as the mean of the alignment matrix as shown in Extended Data Fig. 1c. The Barton-Sternberg algorithm was terminated after four rounds as the s.d. over time remained stable after two rounds (Extended Data Fig. 1e). The mitotic standard time was defined at a temporal resolution of 15 s by subsampling the averaged timeline. To find transitions in the mitotic standard time, the second derivative of the model at each time point for each feature dimension was calculated from

$$x_t'' = |(x_{t-} - x_t) - (x_t - x_{t+})|$$

in which $x_{t-} = \sum_{i=t-18}^{t-1} x_i/18$ and $x_{t+} = \sum_{i=t+1}^{t+18} x_i/18$.

Peaks above a pre-defined threshold were selected across all dimensions as transitions (Extended Data Fig. 2a). In the later part of the model, in which the values of the second derivatives were generally low, small peaks were selected as additional transitions such that no stage between two transitions lasted longer than 12 min. Furthermore, transitions with lower values were deleted to ensure a minimum duration of 1.5 min for each stage (Extended Data Fig. 2b, c).

This approach provides an objective way to discretize the mitotic standard time, which depends on the sampling and the number of cells used. Varying the number of cells sampled from our data identified between 19 and 21 stages with a median set of 20 mitotic stages, which we therefore used throughout the study. To check that these mitotic standard stages were biologically relevant, we automatically selected the 3D image stack closest to the average feature values of each stage. Although the images picked in this way are from different cells, their automatically assigned sequential order reconstitutes a virtual mitosis with an error-free chronology (Extended Data Fig. 2d), in which all classically known mitotic transitions such as NEBD (between stage 2 and 3), and anaphase onset (between stage 11 and 12), were correctly identified. Moreover, the method could identify previously hard-to-define stages such as the first formation of the metaphase plate in late prometaphase (between stage 7 and 8), and could differentiate between the different anaphase and telophase stages (stage 12 to 17). In addition, the kinetics of chromosome condensation is consistent with previous reports in different cell types^{15,35} suggesting that the method could be applied to standardize the mitotic time in other cell types. To test this, we acquired a 4D image dataset consisting of 43 U2OS cell divisions using the same imaging and landmarks approach. The same computational pipeline could indeed generate a mitotic standard time and mitotic standard stages for this cell line (Extended Data Fig. 3).

Modelling of the canonical cell. To support spatial averaging, all cells assigned to the same standard mitotic stage were registered into a common reference coordinate system to give them the same location and orientation. To this end, a virtual coordinate system was defined with its origin at the centre of a volume chosen to be large enough to accommodate all cells after registration. Landmarks (that is, the cell boundary and chromosomal volumes) were then registered to the virtual coordinate system by applying a transformation function involving translation and rotation in 3D. This transformation function was estimated such that the predicted cell axis was aligned with the x axis in the virtual coordinate system. This transformation was applied to both landmarks to preserve their interrelationship in the registered image stacks as shown in Extended Data Fig. 4a, b. Bicubic interpolation was used when applying the transformation⁴².

Registered landmarks were subsequently represented using a cylindrical coordinate system that transforms 3D coordinates into radial distances providing greater flexibility in shape analysis. To this end, we converted landmarks in each plane along the z axis and along the predicted cell axis to polar coordinates in which object boundaries are represented by their radial distances from the object centroid (Extended Data Fig. 4c). As the centroids were aligned on the z axis, the cylindrical representation was formed by concatenating into a vector the polar representations for all planes (Extended Data Fig. 4d). After chromosome segregation, two separate cylindrical representations were used to encode each of the two daughter nuclei.

In this case, the cylindrical axis of each chromosome passes through the centroid of that chromosomal volume.

The standard mitotic space represented by the averaged landmarks was computed in three steps. In the first step, the cylindrical coordinate vectors were averaged separately for each landmark across all cells within each standard mitotic stage (Extended Data Fig. 4e). The average vectors were then transformed back to a Cartesian coordinate system from which binary image stacks were generated. In a second step, to reconstruct the landmarks, the average volume of each landmark was obtained by combining two binary image stacks: one obtained using the z axis as the cylinder axis and the other using the cell axis as cylinder axis (Extended Data Fig. 4f). This combination involved first taking the intersection between the two binary images and then extending it until the average volume of all the cells belonging to the mitotic stage being processed was reached (Extended Data Fig. 4f). Because multiple frames of a cell could be assigned to the same mitotic standard stage, cells could have unequal contributions to each stage with some cells represented more than others at a given stage. To ensure uniform contribution from each cell towards the average mitotic space, in the third step—for each given mitotic stage and for each cell—frame that was most similar to the average shape obtained in step two was selected. These selected cells were then used to re-compute the average shape of the corresponding mitotic stage as described above. This final average shape was also used to calculate the s.d. of all cells in the same mitotic stage. Average mitotic space and standard deviation were generated for all the stages (7–20) for each of the landmarks (examples in Extended Data Fig. 4g, h).

Generation of the protein density map. Standard mitotic spaces were used as the reference to register and integrate protein distributions from many different cells to generate protein density maps (Fig. 3a). All calibrated protein concentration maps having the same protein in a given mitotic stage were registered first to the corresponding standard mitotic space using the predicted cell-division axis. This transformed all individual protein image stacks to the same coordinate system. Bicubic interpolation⁴² was used during the rotation. Registered image stacks were then accumulated in the standard mitotic space. Pixels outside the segmented cell region and mapped outside the standard mitotic space were discarded. A protein density map was then created by averaging the accumulated intensities in the standard mitotic space (Fig. 3a). Density maps of all proteins for mitotic stages 7–20 were estimated in the same way and can be explored on http://www.mitocheck.org/mitotic_cell_atlas.

Feature extraction of images. The protein z -stack concentration map was processed using a Gaussian filter (Matlab `smooth3` function with a kernel of size [3 3 1] and s.d. 0.65) and a maximum projection along the z -axis and normalization to the theoretical saturation intensity. SURF interest points¹⁸ were then detected on the image resized to a 0.063- μm resolution using three octaves each including four Haar wavelet filters at different sizes from 9-by-9 to 99-by-99 pixels, ranging from about 0.5 μm to more than 6 μm . Interest points were further selected such that most of the protein signals were counted in one of the interest points. Each of these interest points was then described by a numerical vector quantifying features in the following four categories: locations relative to the landmarks (four features), correlation to the H2B signal or the predicted midplane/midbody volume according to the localization of the interest point (one feature), flattened soft spin image features⁴³ describing the intensity distribution within an interest point (30 features), and summarized uniform local binary patterns⁴⁴ describing the orientation of the signal (4 features).

Five per cent of the cells were randomly selected and their interest points were used to construct a training set for identifying clusters of interest points with similar features (Extended Data Fig. 6a). All training interest points were first separated into 16 clusters by their localization feature and the quarter-level of their metric values. Interest points in clusters with a sufficient size were then further clustered on the basis of correlation features separated by pre-defined thresholds and a `dbscan`⁴⁵ clustering for each sub-cluster in the reduced feature space covering 85% of the variance according to a principal component analysis⁴⁶ on the uniform local binary patterns and spin image features⁴³. The final clustering step was performed only for the clusters with the highest-contrast value in their location category on the basis of the spin image features in which the interest points were separated into homogeneous bright, structured bright and dim clusters by a pre-defined threshold. The total number of clusters was not deterministic as the training set was randomly generated, but eight rounds of clustering yielded between 87 and 100 clusters and a run with 100 clusters was used for further analysis. Interest points in the same cluster share similar textures (Extended Data Fig. 6b). All interest points in the remaining images in the dataset were then each assigned to one cluster. The total intensity within each interest point was then counted and the fraction of intensities recorded in each interest point cluster was calculated for each cell so that each image was represented by a 100-dimensional feature vector with a sum of one.

Non-negative tensor factorization. Each protein was represented at each mitotic stage by the average of all its vectors present at that stage. Owing to the binning of

consecutive imaging time points, a cell can be represented by several vectors at a given mitotic stage. These duplicates were replaced by their mean, resulting in each cell being represented by only one vector per mitotic stage. The resulting dataset is a three-dimensional tensor X of 28 proteins \times 100 features \times 20 mitotic stages. We view canonical subcellular localizations as latent features of the data, that is, we assume that, at any time point, the observed vector for a protein was generated by a combination of the different canonical subcellular localizations the protein occupied at this stage. A protein vector x can then be expressed as the product of a subcellular localization membership vector z and a matrix A of canonical subcellular localization features: $x = zA$. Therefore, we wish to model our data tensor X such that for each frontal (temporal) slice X_t ,

$$X_t = Z_t A + E_t (t = 1, 2, \dots, 20)$$

in which Z_t is a matrix whose rows are localization membership vectors and E_t is a matrix containing the errors.

Given that all feature values are non-negative, a possible solution for each time point can be found by non-negative matrix factorization of individual matrices X_t ⁴⁷. However, processing time points independently results in loss of information with the undesirable effect that different canonical localizations are learned for different time points. Simultaneous non-negative factorization of a set of matrices is a special form of non-negative tensor factorization that can be reduced to a standard non-negative matrix factorization using column-wise unfolding of the data tensor X ⁴⁸:

$$X = ZA + E$$

in which X is formed by vertically stacking the X_t matrices and Z is formed by the correspondingly stacked Z_t matrices and E contains the errors. Z and A are then found using multiplicative updates⁴⁷ to minimize the objective function $\|X - ZA\|$ where $\|\cdot\|$ indicates the Frobenius norm. As a final step, the rows of Z are normalized to sum one. Values in Z can be interpreted as fractions of the amount of protein (captured by the features) present at each canonical localization.

The method requires choosing the number k of canonical subcellular localizations we want to represent our data with. There is no good strategy for finding this number *a priori* because increasing k corresponds to a higher resolution of the localization description. For example, a low k results in lumping all chromatin proteins together whereas a higher k resolves kinetochore proteins from other chromatin proteins. Thus the optimal number of subcellular localizations is partly subjective, depending on the level of granularity desired. However, we can use heuristics to help guide the choice of k . If the number of selected canonical localizations is too low, many proteins will share the same temporal profile, that is, their corresponding vectors in Z_t will be highly similar for all time points. As more canonical localizations are added, we can expect more proteins to resolve into distinct profiles, that is, the similarity between their corresponding vectors will decrease until eventually adding more canonical localizations will not improve resolution and similarity will stop decreasing. Similarity between vectors across time points can be measured using Tucker's congruence coefficient⁴⁹. Therefore, for each value of k from 2 to 25, we plot the fraction of Tucker's congruence coefficient values above 0.6. The value of k for which the fraction of highly similar proteins reaches a low-value plateau indicates that there are enough canonical localizations to describe each protein individually and therefore this value of k represents an upper bound on the number of canonical localizations. According to this procedure, k was set to seven for the current data. Because the non-negative matrix factorization algorithm can converge to a local minimum of the objective function, ten runs with random initialization of the matrices were performed and the run with lowest value of the objective function was kept. A flattened representation of the resulting tensor can be obtained by assigning a different colour to each cluster and plotting each protein distribution as a bar chart in which the height of each colour band at each time point is proportional to the fraction of the protein amount in the corresponding cluster (Extended Data Fig. 7). A dynamic multigraph can be derived from the cluster memberships as follows: first an edge type is defined for each cluster. If two genes share a cluster at a given time point, then an edge of that type is added between them at that time point. The edge weight is set to the product of the linked genes fractions in the corresponding cluster and can be loosely interpreted as a probability of interaction. For visualization, only edges with a weight greater than an arbitrary threshold (here set to 0.3) were kept (Fig. 4b).

Analysis of protein localization kinetics using supervised annotation. A multivariate linear regression model with a multivariate Gaussian response was trained with an elastic net regularization and non-negativity constraints on the coefficients with the feature vectors described above as predictors and localization vectors as response. The response vectors were defined using cells with tagged proteins known to be specific markers of unique subcellular compartments (Extended Data Table 1) as follows: for each of the marker proteins, the fraction of total intensity in the foreground was determined by Otsu thresholding of the 3D image

stack and the corresponding protein amount assigned to the compartment with the complement assigned to cytoplasm. Each cell is thus represented by a seven-dimensional response vector containing the fraction of the tagged protein in the following compartments: chromatin, kinetochore, centrosome, spindle, midbody, nuclear envelope and cytoplasm. To deal with the compositional nature of this data, all features and response vectors are transformed using the additive log-ratio transformation^{50,51} with the inverse hyperbolic sine function as a generalized logarithm to handle occurrences of 0. The model with the best fit using fivefold cross-validation was selected.

The predictions from the model were transformed back to proportions using the inverse of the log-ratio transformation then multiplied by the total number of proteins to predict the absolute number of molecules in each mitotic subcellular structure for each image. Predictions were then smoothed using local polynomial regression fitting.

To compute the anaphase dissociation kinetics for each kinetochore protein (Extended Data Fig. 8e), we fitted each curve between 30 min and 42 min mitotic standard time (late metaphase to telophase) with a four parameter sigmoidal decay function:

$$y = \frac{a-d}{(1 + bc^{-\text{time}})} + d$$

for which the first and second derivatives were analytically calculated. The time of disassembly was defined as the point at which the second derivative is equal to 0 (inflection point of the curve). The disassembly rate was computed as the minimum value of the first derivative in the time interval.

Statistics and reproducibility. For each protein, the number of cells and number of experiments that were run to collect them is reported in Supplementary Table 1. Unless stated otherwise, all cells for a given protein were used in the reported analyses.

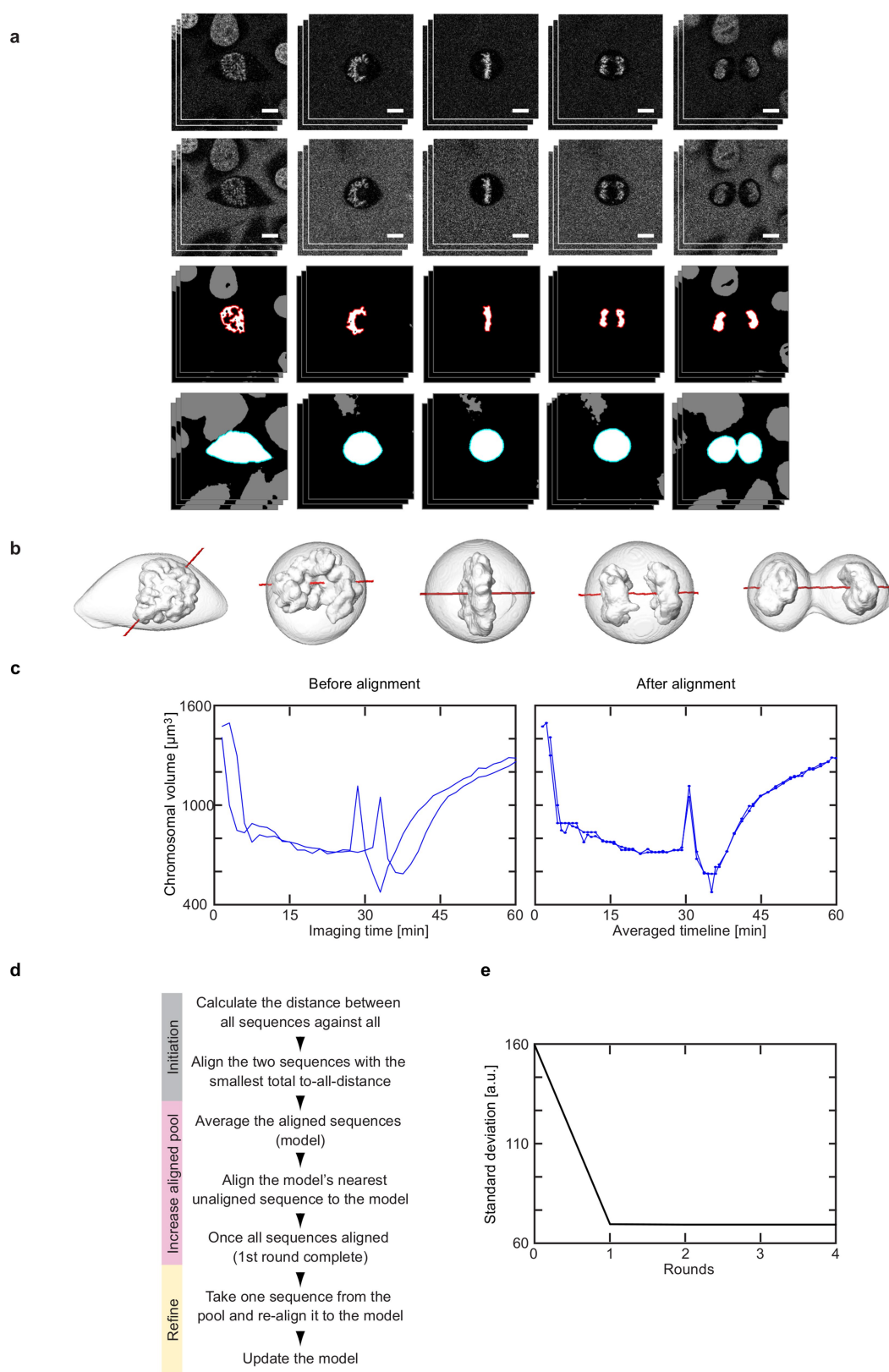
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. All source code is accessible on EMBL's GitLab instance: https://git.embl.de/grp-ellenberg/mitotic_cell_atlas and can be downloaded or cloned using the command 'git clone https://git.embl.de/grp-ellenberg/mitotic_cell_atlas'. or on the project website at http://www.mitocheck.org/mitotic_cell_atlas/downloads/v1.0.1/mitotic_cell_atlas_v1.0.1_src.zip. Instructions to run the code are provided as a README file together with the source code. An example dataset to run and test the source code can be downloaded from http://www.mitocheck.org/mitotic_cell_atlas/downloads/v1.0.1/mitotic_cell_atlas_v1.0.1_exampledata.zip.

Data availability. All images processed in this study including original images, concentration maps, segmentation mask for both cellular and chromosomal volume and concentration maps are available in the Image Data Resource (<http://idr.openmicroscopy.org>⁵¹) under DOI: 10.17867/10000112. Further data and code are available as follows: all images are also available for download on the mitotic cell atlas website http://www.mitocheck.org/mitotic_cell_atlas/downloads/v1.0.1/mitotic_cell_atlas_v1.0.1_fulldata.zip (~0.5 TB). The data supporting the spatio-temporal mitotic cell model and the analysis is available from the mitotic cell atlas website (http://www.mitocheck.org/mitotic_cell_atlas/downloads/v1.0.1) and contains: i) segmentation masks for the landmarks (that is, cell boundary and chromosome mass(es)) as TIFF files (directory 'mitotic_cell_model/binary_masks') and snapshots of the 3D rendering of each of the spatial models in VRML and TIFF formats (directory 'mitotic_cell_model/snapshots'). ii) Two movies (orthogonal and oblique views) created from 3D reconstructed average landmarks (cell boundary and chromosome mass(es); directory 'mitotic_cell_model/movies'). iii) Average concentrations of each protein at individual mitotic stages as mat files, TIFF stacks, and tab-delimited text files (directory 'protein_distributions'). iv) Feature data used for the analysis (to produce Fig. 4, Extended Data Figs. 7, 8d, e, 9)

in a tab-delimited text file (file 'cell_features.txt'). This file can be used directly as input to the notebooks available in the code repository. This file also contains the mitotic standard time and stage assigned to each cell image. v) Canonical localization data (file 'canonical_mitotic_clusters.h5'). vi) Dynamic graph (file 'dynamic_graph_adjacency_matrices.h5').

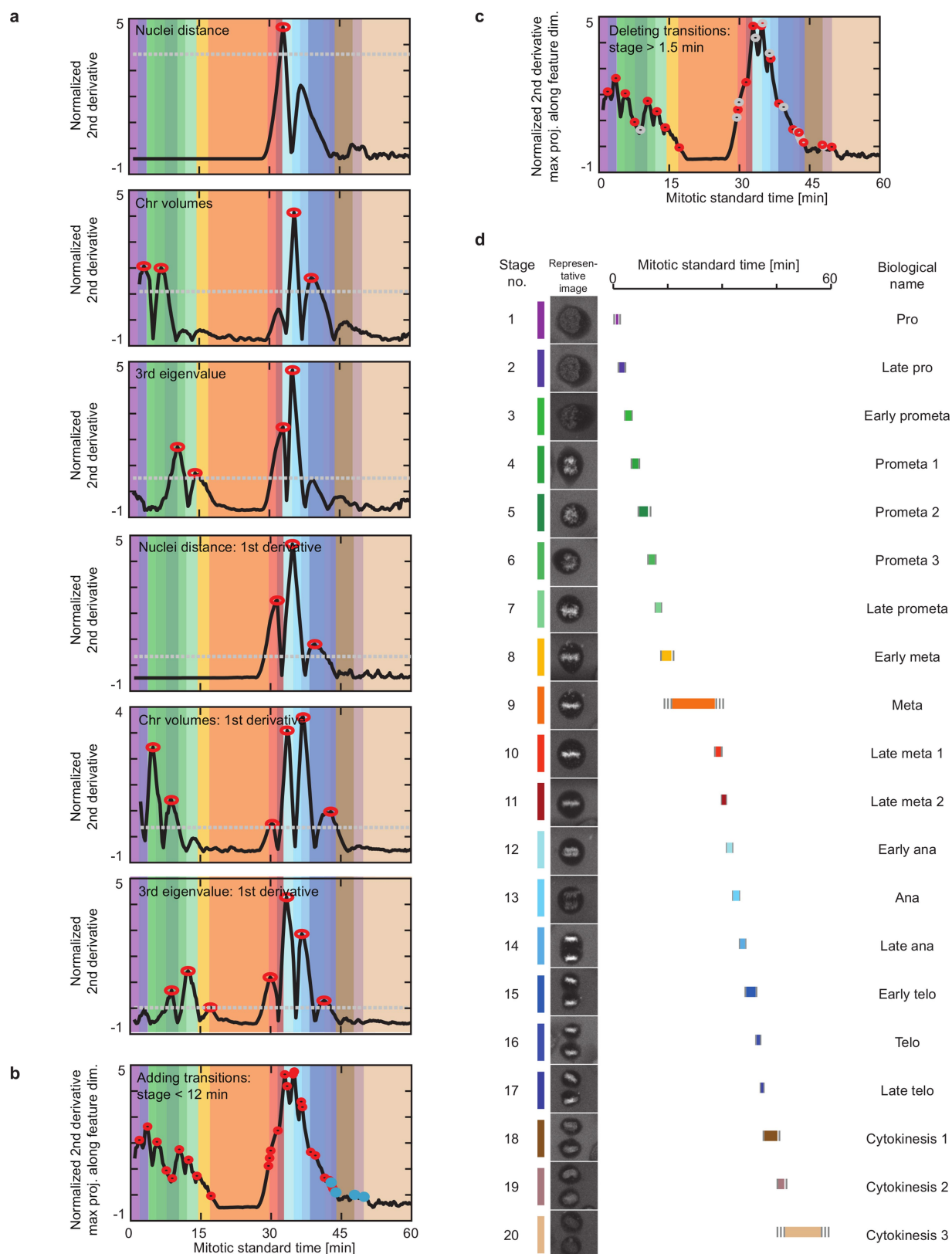
25. Maliga, Z. et al. A genomic toolkit to investigate kinesin and myosin motor function in cells. *Nat. Cell Biol.* **15**, 325–334 (2013).
26. Hutchins, J. R. et al. Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science*. **328**, 593–599 (2010).
27. Kunitoku, N. et al. CENP-A phosphorylation by Aurora-A in prophase is required for enrichment of Aurora-B at inner centromeres and for kinetochore function. *Dev. Cell* **5**, 853–864 (2003).
28. Mahen, R. et al. Comparative assessment of fluorescent transgene methods for quantitative imaging in human cells. *Mol. Biol. Cell* **25**, 3610–3618 (2014).
29. Walther, N. et al. A quantitative map of human Condensins provides new insights into mitotic chromosome architecture. *J. Cell Biol.* **217**, 2309–2328 (2018).
30. Ladurner, R. et al. Sororin actively maintains sister chromatid cohesion. *EMBO J.* **35**, 635–653 (2016).
31. Davidson, I. F. et al. Rapid movement and transcriptional re-localization of human cohesin on DNA. *EMBO J.* **35**, 2671–2685 (2016).
32. Neumann, B. et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464**, 721–727 (2010).
33. Bancaud, A. et al. Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin. *EMBO J.* **28**, 3785–3798 (2009).
34. Li, C. H. & Lee, C. K. Minimum cross entropy thresholding. *Pattern Recognit.* **26**, 617–625 (1993).
35. Hériché, J.-K. et al. Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation. *Mol. Biol. Cell* **25**, 2522–2536 (2014).
36. Otsu, N. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
37. Meyer, F. Topographic distance and watershed lines. *Signal Processing* **38**, 113–125 (1994).
38. Wachsmuth, M. et al. High-throughput fluorescence correlation spectroscopy enables analysis of proteome dynamics in living cells. *Nat. Biotechnol.* **33**, 384–389 (2015).
39. Ori, A. et al. Cell type-specific nuclear pores: a case in point for context-dependent stoichiometry of molecular machines. *Mol. Syst. Biol.* **9**, 648 (2013).
40. Barton, G. J. & Sternberg, M. J. E. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J. Mol. Biol.* **198**, 327–337 (1987).
41. ten Holt, G., Reinders, M. & Hendriks, E. Multi-dimensional dynamic time warping for gesture recognition. In *Proc. 13th Annual Conference of the Advanced School for Computing and Imaging* (2007).
42. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust.* **29**, 1153–1160 (1981).
43. Lazebnik, S., Schmid, C. & Ponce, J. A sparse texture representation using local affine regions. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1265–1278 (2005).
44. Heikkilä, M. & Pietikäinen, M. A texture-based method for modeling the background and detecting moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 657–662 (2006).
45. Tran, T. N., Drab, K. & Daszykowski, M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemom. Intell. Lab. Syst.* **120**, 92–96 (2013).
46. Jolliffe, I. T. *Principal Component Analysis* (Springer, New York, 2002).
47. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
48. Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S.-I. *Nonnegative Matrix and Tensor Factorizations* (John Wiley & Sons, Chichester, 2009).
49. Tucker, L. R. *A Method for Synthesis of Factor Analysis Studies* (Educational Testing Service, Princeton, 1951).
50. Aitchison, J. *The Statistical Analysis of Compositional Data* (Blackburn Press, Caldwell, 2003).
51. Williams, E. et al. Image Data Resource: a bioimage data integration and publication platform. *Nat. Methods* **14**, 775–781 (2017).



Extended Data Fig. 1 | Segmentation and time alignment.

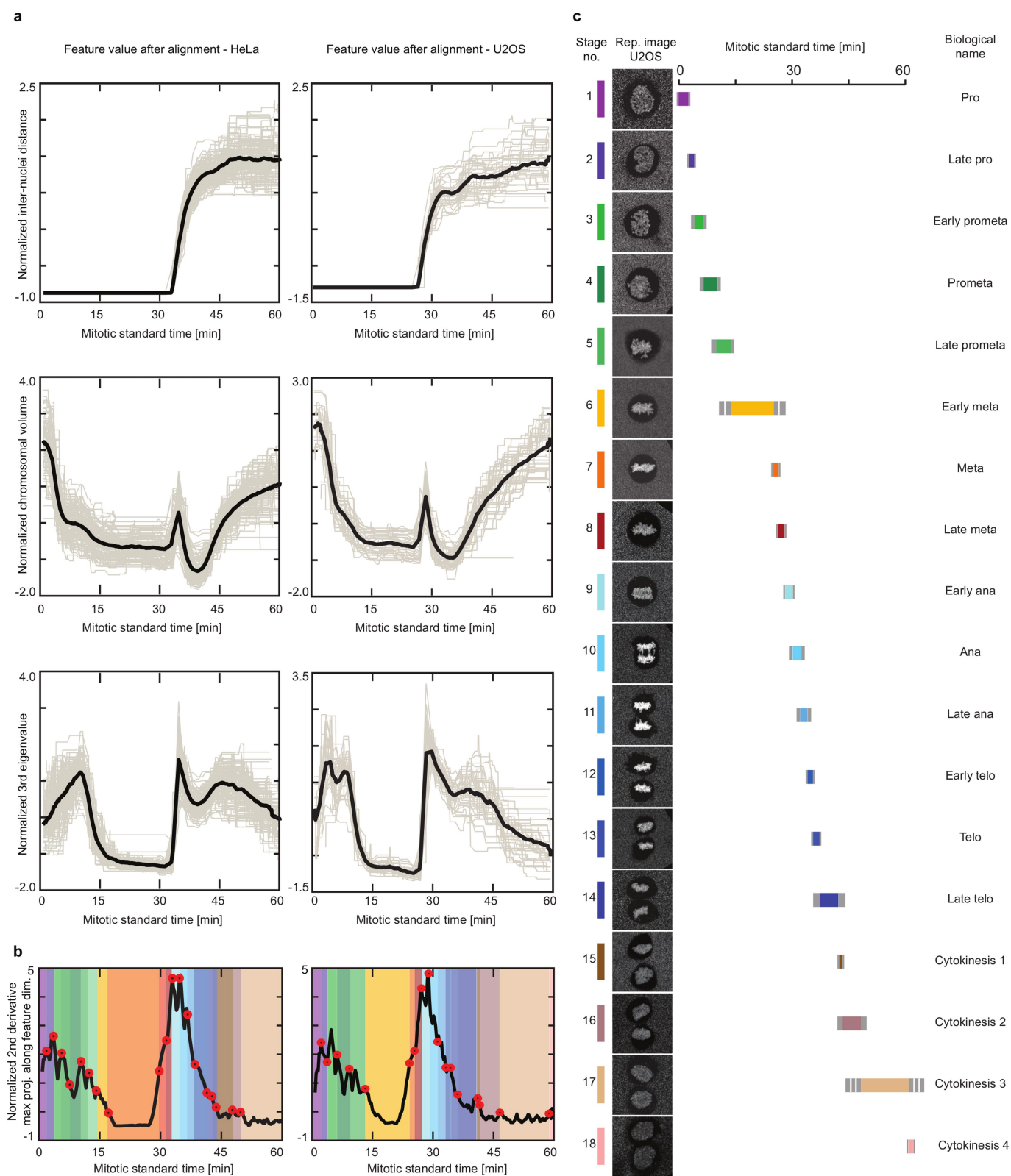
a, b, Segmentation and 3D reconstruction of landmarks. **a**, Single x - y plane image in mCherry (587–621 nm, first row) and DY481XL (622–695 nm, second row) detection channels. Third row: detected chromatin markers in which boundaries of the chromosomal volume of interest are marked in red. Fourth row: output of watershed transform on ratio image in which the boundary of the detected cell of interest is marked in green. Scale bar, 10 μ m. **b**, Reconstruction of cell and chromosomal surfaces in

3D (grey) and the predicted division axis (red). **c–e**, Generating the mitotic standard time model. **c**, Dynamic time warping is used to align a pair of time-resolved sequences. **d**, Modified Barton–Sternberg algorithm to align 132 sequences. **e**, The cumulative s.d. of a single feature after each iteration of the algorithm. It remains nearly constant after the second round indicating that at termination (fourth round) a stable time alignment was achieved. This has been repeated 10 times and similar alignment results are obtained when the number of cells is more than 50.



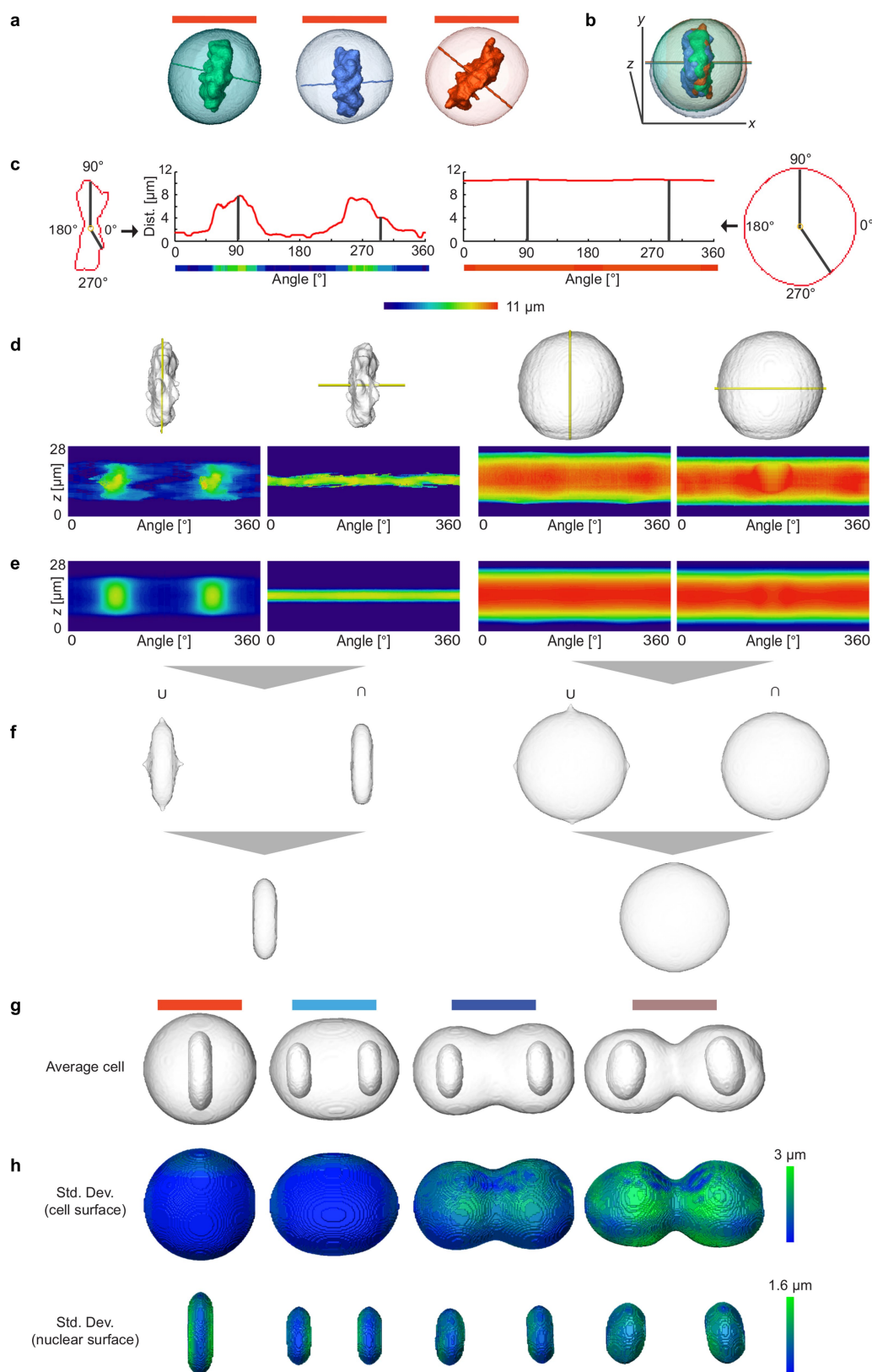
Extended Data Fig. 2 | Detection of mitotic standard stages. a, Detection of major mitotic transitions of the mitotic standard time. Peaks in the second derivatives (red circles) above a pre-defined threshold (grey lines) were detected in all feature dimensions as mitotic transitions. **b**, Additional smaller peaks (blue dots) were detected to ensure a

maximum duration of 12 min for each standard stage. **c**, Transitions were deleted (grey circles) such that all stages had a minimal duration of 1.5 min. **d**, The standard mitotic cell was represented by the cell closest to the mean of each stage. Each mitotic stage was assigned duration (coloured line), its duration s.d. (grey line) and a biological annotation.



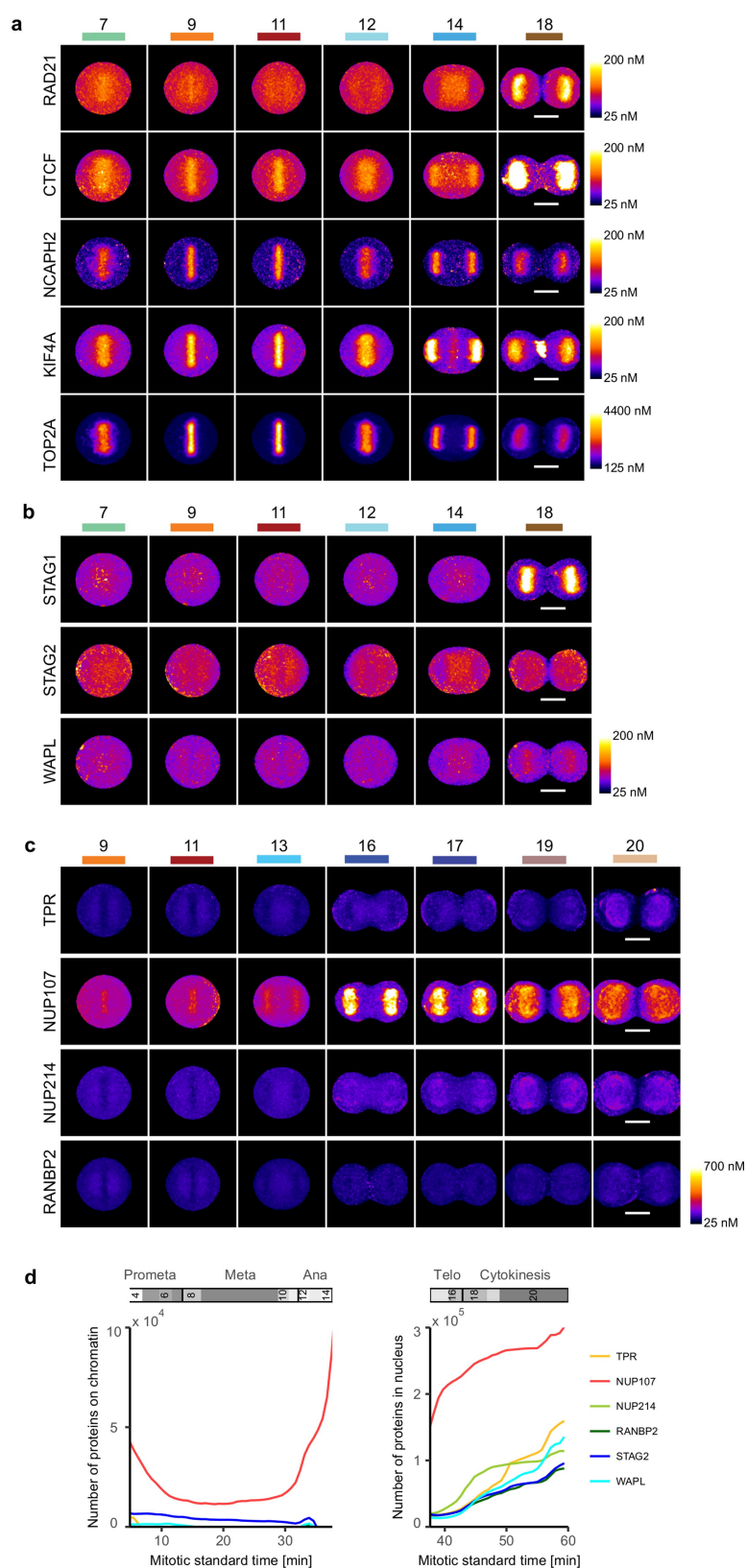
Extended Data Fig. 3 | Comparison between mitotic standard time for HeLa Kyoto and U2OS cells. a, Features used for generating the mitotic standard time model after alignment for HeLa Kyoto cells (left) and U2OS cells (right). Grey line, normalized feature value over time of individual cells; black line, mean. **b**, Mitotic standard time transitions for HeLa cells

(left) and U2OS cells (right). **c**, Standard mitotic U2OS cell represented by the cell closest to the average of each mitotic standard stage. Each mitotic stage was assigned duration (coloured line), its duration s.d. (grey line) and a biological annotation.



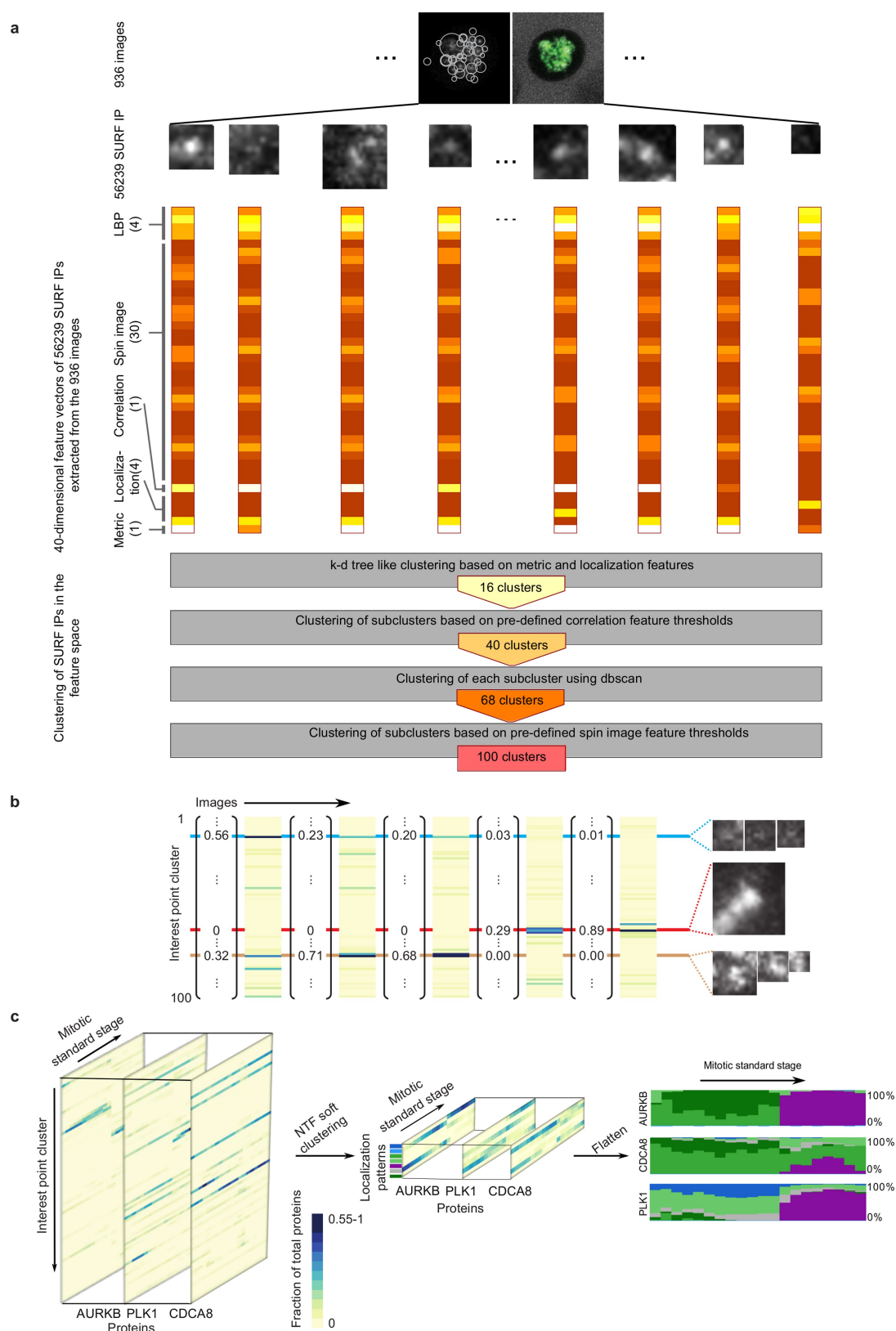
Extended Data Fig. 4 | Generation of spatial model for standard mitotic stages by combining two cylindrical representations. **a, b**, Examples of cells in mitotic stage 10 (**a**) were registered using the predicted cell division axis as shown in **b**. **c**, Transformation between Cartesian and cylindrical coordinate systems. **d**, Example cellular and chromosomal surfaces (grey) were transformed into the cylindrical coordinate system using two cylindrical axes (z-axis or predicted division axis) marked in yellow. **e**, Average cellular and chromosomal surfaces in cylindrical coordinate

systems. **f**, Union (U) and intersection (∩) of the averaged landmarks volumes represented in the Cartesian coordinate system that were then combined to generate final cellular and chromosomal surfaces shown in the first image in **g**. **g**, By averaging a large number of cells, models were generated for all mitotic standard stages with symmetrical geometries and example stages 10, 14, 16 and 19 are shown. **h**, The spatial variation of the mitotic standard spaces shown in **g**.



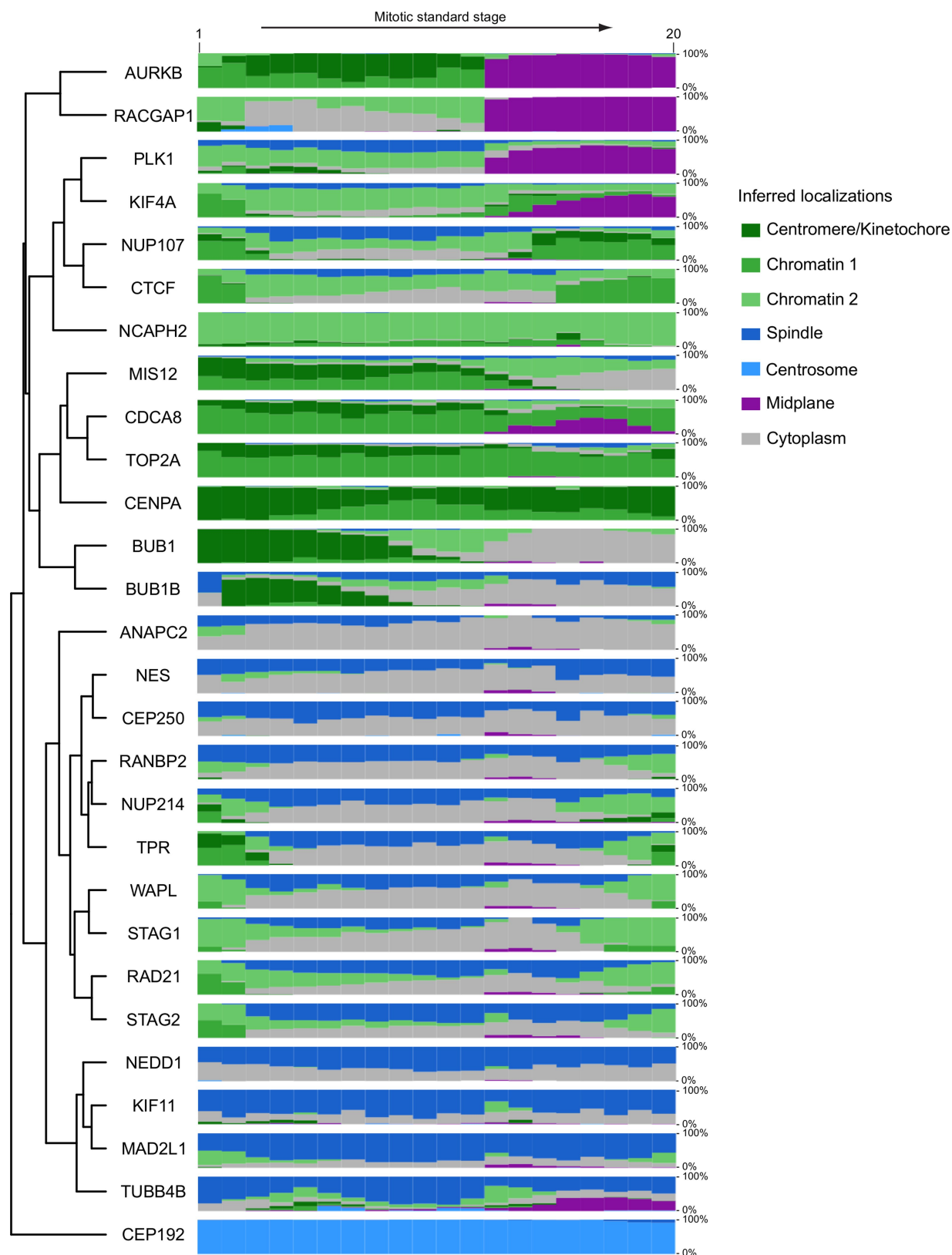
Extended Data Fig. 5 | Chromatin organizers and NUPs localization.
a–c, Maximal intensity projection from the mitotic standard model at selected stages. Scale bars, 10 μm . **a**, Chromatin organizers RAD21, CTCF, NCAPH2, KIF4A and TOP2A present on chromatin during mitosis.
b, Chromatin organizers STAG1, STAG2 and WAPL with weak binding

to chromatin during mitosis. **c**, Four NUPs at selected standard mitotic stages. **d**, NUPs localization as function of mitotic standard time. The curves for STAG2 and WAPL are shown as a reference and are identical to the data from Fig. 3c.



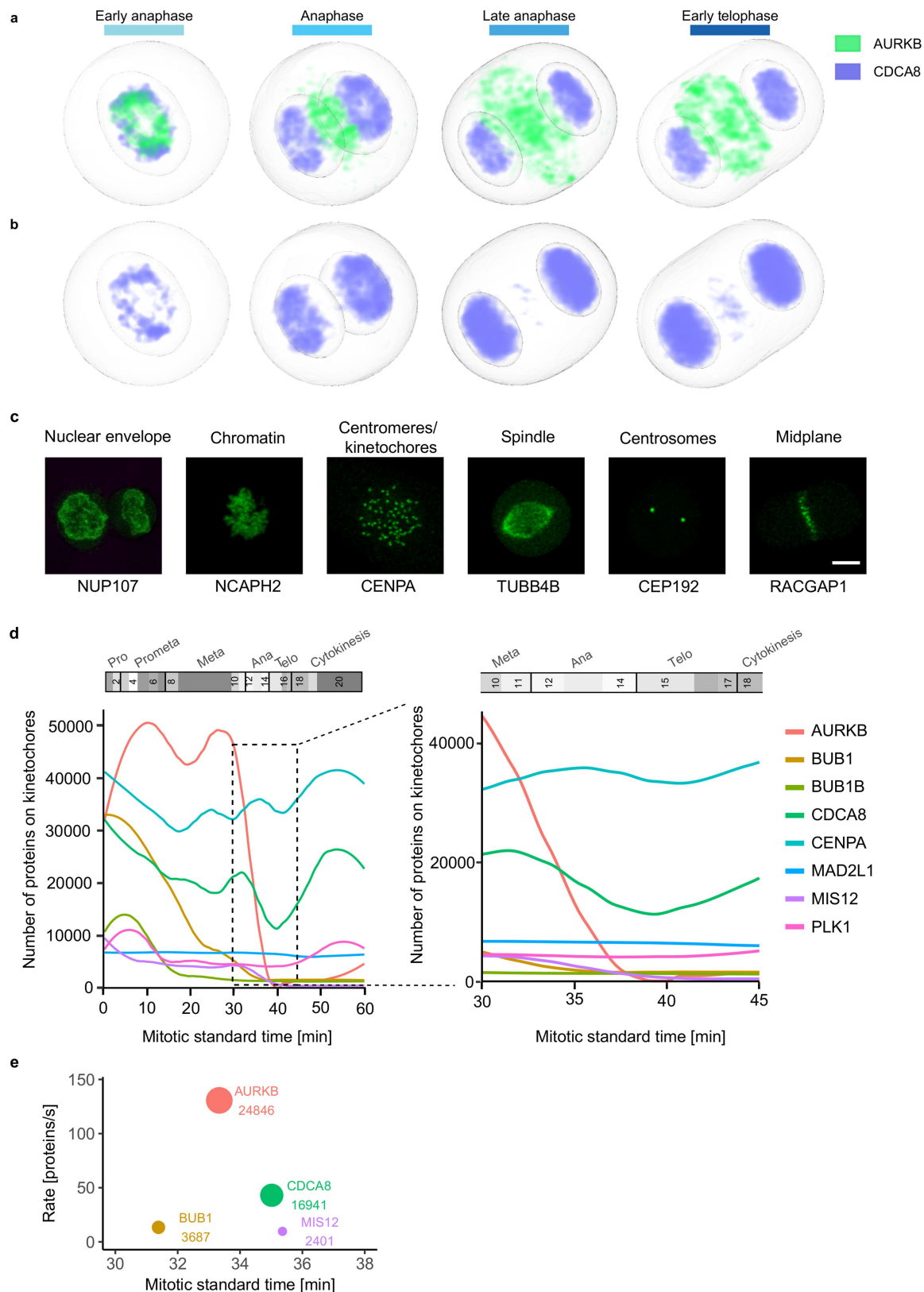
Extended Data Fig. 6 | Interest point clusters and dynamic protein localization. **a**, Pipeline for the definition of interest point clusters using a subset of the data. Images (936, corresponding to 5% of the entire dataset) were randomly selected from the dataset to construct a pool of interest points. Each interest point was numerically described with a 40 dimensional feature vector encoding the intensity distribution, localization and contrasts to the interest point neighbourhood. Combining *k*-d tree-like and thresholding-based clustering with density-based clustering, the interest points were grouped into 100 clusters. **b**, The remaining interest

points of the dataset were then assigned to the identified clusters. Thus each image was represented as the distribution of intensity in each of the 100 interest point clusters. **c**, Non-negative factorization of the data tensor of proteins \times features \times mitotic stages (left panel) produced a non-negative tensor of reduced dimension (middle) for which entries can be interpreted as the fraction of protein belonging to each cluster over time (right, each cluster is represented by a different colour and the height of a coloured bar at a given mitotic stage represents the fraction of the protein in the corresponding cluster at this stage).



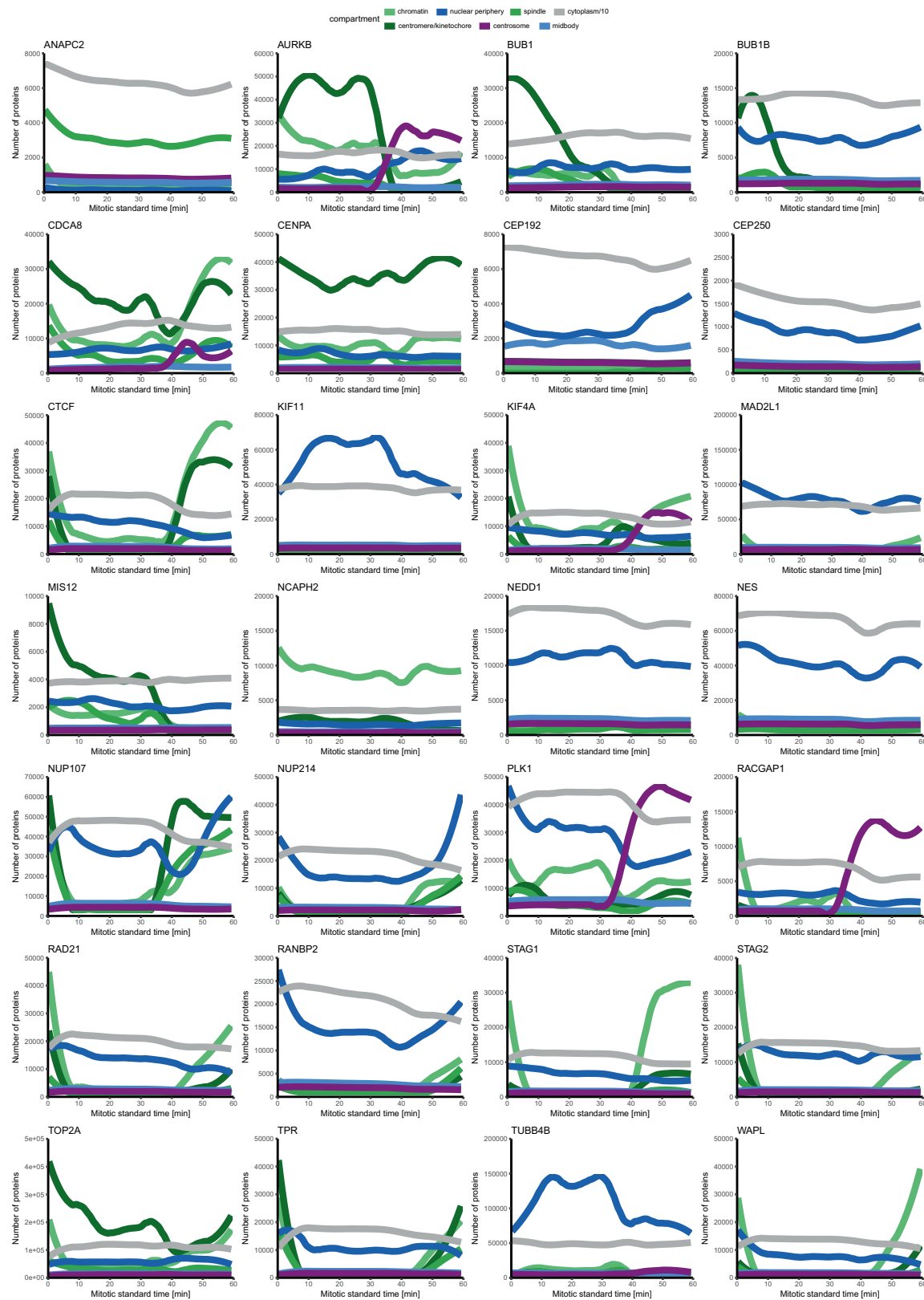
Extended Data Fig. 7 | Quantitative evolution of protein subcellular localizations inferred from non-negative tensor factorization of the proteins \times features \times time tensor. Each subcellular localization cluster was assigned a different colour and named using known information on

proteins belonging to that cluster. The height of each colour band at each time point is proportional to the fraction of the protein amount in the corresponding cluster at that time point. Genes were grouped by complete linkage clustering with optimal leaf ordering.



Extended Data Fig. 8 | Mitotic standard model and supervised classification to investigate the dynamic localization of kinetochore proteins. **a, b**, Concentration maps of chromosome passenger proteins AURKB and CDCA8 in anaphase and early telophase. **a**, AURKB concentrates in an outer ring and a central disk. Most of CDCA8 remains on chromatin, and after AURKB has already relocalized—between late anaphase and early telophase—only a small CDCA8 fraction colocalizes with AURKB in the central disk. **b**, Colour displaying CDCA8 was adapted to make its localization in the central disk visible. **c–e**, Analysing

sub-cellular (dis)assembly kinetics using a supervised approach. **c**, Example of maximally z-projected images of marker proteins for the selected subcellular compartments used for the supervised approach. Scale bar, 10 μm . **d**, Kinetics of kinetochore disassembly. The predicted number of molecules localized on kinetochore and centromeres are plotted for eight proteins in the mitotic standard time (left) and zoomed in for anaphase (right). **e**, Order and rate of protein removal from the kinetochore during anaphase. The annotation and circle diameter indicate the number of molecules at the estimated average time of dissociation.



Extended Data Fig. 9 | Prediction of protein molecule numbers on major mitotic subcellular structures using the supervised approach. The colour scheme is adjusted to the most similar cluster identified using

non-negative tensor factorization (Extended Data Fig. 7). Cytoplasm values are divided by ten.

Extended Data Table 1 | Reference structures for supervised model

Localization	Gene	Mitotic standard stages
Nuclear envelope	NUP107	15-20
Chromatin	NCAPH2	1-20
Kinetochores	CENPA	1-20
Centrosomes	CEP192	1-20
Spindle	TUBB4B	4-20
Midbody	RACGAP1	12-20

In vivo CRISPR editing with no detectable genome-wide off-target mutations

Pinar Akcakaya^{1,13}, Maggie L. Bobbin^{2,3,4,13}, Jimmy A. Guo^{2,3}, Jose Malagon-Lopez^{2,3,4}, Kendell Clement^{2,3,4}, Sara P. Garcia², Mick D. Fellows⁵, Michelle J. Porritt¹, Mike A. Firth⁶, Alba Carreras^{1,9}, Tania Baccaga^{1,10}, Frank Seeliger⁷, Mikael Bjursell¹, Shengdar Q. Tsai^{2,3,4,11}, Nhu T. Nguyen^{2,3}, Roberto Nitsch⁸, Lorenz M. Mayr^{1,12}, Luca Pinello^{2,4}, Mohammad Bohlooly-Y¹, Martin J. Aryee^{2,4}, Marcello Maresca^{1*} & J. Keith Joung^{2,3,4*}

CRISPR–Cas genome-editing nucleases hold substantial promise for developing human therapeutic applications^{1–6} but identifying unwanted off-target mutations is important for clinical translation⁷. A well-validated method that can reliably identify off-targets in vivo has not been described to date, which means it is currently unclear whether and how frequently these mutations occur. Here we describe ‘verification of in vivo off-targets’ (VIVO), a highly sensitive strategy that can robustly identify the genome-wide off-target effects of CRISPR–Cas nucleases in vivo. We use VIVO and a guide RNA deliberately designed to be promiscuous to show that CRISPR–Cas nucleases can induce substantial off-target mutations in mouse livers in vivo. More importantly, we also use VIVO to show that appropriately designed guide RNAs can direct efficient in vivo editing in mouse livers with no detectable off-target mutations. VIVO provides a general strategy for defining and quantifying the off-target effects of gene-editing nucleases in whole organisms, thereby providing a blueprint to foster the development of therapeutic strategies that use in vivo gene editing.

VIVO consists of two steps (Fig. 1a). In an initial in vitro ‘discovery’ step, a superset of potential off-target cleavage sites for a nuclease is identified using circularization for in vitro reporting of cleavage effects by sequencing (CIRCLE-seq)⁸. This method is highly sensitive, avoids potential confounding effects associated with cell-based assays⁸ and can successfully identify supersets of sites that include bona fide off-targets in cultured human cells⁸. In a second in vivo ‘confirmation’ step, sites identified by CIRCLE-seq are examined for indel mutations in target tissues that have been treated with the nuclease.

To test VIVO, we designed a *Streptococcus pyogenes* Cas9 (hereafter, Cas9) guide RNA (gRNA) targeted to the mouse *Pcsk9* gene that we expected would have a high likelihood of inducing multiple off-target mutations in the mouse genome (Extended Data Fig. 1a, Extended Data Table 1 and Methods). To deliver this ‘promiscuous’ gRNA (gP) and Cas9 to mouse livers in vivo, we infected cohorts of mice with an adenoviral vector that encodes these components or with a negative control vector that encodes GFP and Cas9. Adenoviral vectors have a known biodistribution, which includes efficient delivery to the liver (Extended Data Fig. 2). Two strains of mice were infected: a wild-type C57BL/6N strain (hereafter, wild-type mice) and littermates that contain a single copy of a human *PCSK9* open reading frame knocked into the *Rosa26* locus (hereafter, knock-in mice; A.C. et al., manuscript submitted; Extended Data Fig. 3 and Methods). gP–Cas9 induced stable and efficient mutation of the on-target *Pcsk9* site and reductions in *Pcsk9* protein levels in plasma, in both wild-type and knock-in mice (Extended Data Fig. 1b, c).

Having established the efficacy of gP–Cas9 for on-target *Pcsk9* modification in vivo, we conducted the first screening step of VIVO by performing CIRCLE-seq with gP–Cas9 on liver genomic DNA from wild-type and knock-in mice (Fig. 1a and Methods). We identified many off-target cleavage sites in vitro: 3,107 and 2,663 sites with wild-type and knock-in mice genomes, respectively (Extended Data Fig. 4 and Supplementary Table 1). These sites represent only a small percentage of all sites in the genome that have seven or fewer mismatches relative to the on-target site (Extended Data Fig. 5). There were 2,368 sites that were identified in both mouse genomes and that showed strong concordance ($r^2 = 0.902$) in their CIRCLE-seq read counts (Extended Data Fig. 4), which semi-quantitatively reflect cleavage efficiency⁸. Sites identified in only one or the other genome generally had among the lowest CIRCLE-seq read counts (Extended Data Fig. 4). This is consistent with the possibility that these sites may or may not be detected by chance alone because they lie at the assay limit of detection, as has previously been observed with analogous experiments in human cells in culture⁸. Single nucleotide polymorphisms or indels might also have a role in accounting for a very small subset of these sites (Extended Data Table 2). The 20 sites with the highest CIRCLE-seq read counts all had 3 or fewer mismatches in the spacer sequence (Extended Data Fig. 4 and Supplementary Table 1). Many off-target sites contained protospacer adjacent motif (PAM) mismatches, with NAG as the most prevalent (Extended Data Fig. 4 and Supplementary Table 1)—probably because gP has an unusually high relative number of closely matched sites in the mouse genome with a NAG PAM (Extended Data Table 3), a known alternative PAM for Cas9^{8,9}.

To perform the second step of VIVO, we assessed whether the off-target cleavage sites of gP–Cas9 identified by CIRCLE-seq showed evidence of indels in vivo in the livers of wild-type and knock-in mice treated with gP–Cas9 and control GFP–Cas9 adenoviral vectors. Because of the very large number of sites identified by CIRCLE-seq, we performed targeted amplicon sequencing using liver genomic DNA from mice euthanized on day four and on week three after infection on only the following subsets of sites: the *Pcsk9* on-target site, 11 ‘class I’ sites with the highest CIRCLE-seq read counts (containing 1–3 mismatches relative to the on-target site), 17 ‘class II’ sites with moderate CIRCLE-seq read counts (containing 2–4 mismatches) and 17 ‘class III’ sites with lower CIRCLE-seq read counts (containing 1–6 mismatches) (Fig. 1b and Extended Data Fig. 4). The *Pcsk9* on-target site was efficiently mutagenized and remained stably mutated in both wild-type and knock-in mice (mean indel frequencies ranging from about 23 to 30%) (Fig. 1b and Supplementary Table 2). Of the

¹Discovery Biology, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden. ²Molecular Pathology Unit and Center for Cancer Research, Massachusetts General Hospital, Charlestown, MA, USA. ³Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA, USA. ⁴Department of Pathology, Harvard Medical School, Boston, MA, USA. ⁵Advanced Medicines Safety, Drug Safety and Metabolism, IMED Biotech Unit, AstraZeneca, Cambridge, UK. ⁶Quantitative Biology, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Cambridge, UK. ⁷Pathology Science, Drug Safety and Metabolism, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden. ⁸Advanced Medicines Safety, Drug Safety and Metabolism, IMED Biotech Unit, AstraZeneca, Gothenburg, Sweden. ⁹Present address: Wallenberg Laboratory and Sahlgrenska Center for Cardiovascular and Metabolic Research, Department of Molecular and Clinical Medicine, University of Gothenburg, Gothenburg, Sweden. ¹⁰Present address: San Raffaele Telethon Institute for Gene Therapy, IRCCS San Raffaele Scientific Institute, Milan, Italy. ¹¹Present address: Department of Hematology, St. Jude Children’s Research Hospital, Memphis, TN, USA. ¹²Present address: GE Healthcare Life Sciences, The Grove Centre, Amersham, UK. ¹³These authors contributed equally: Pinar Akcakaya, Maggie L. Bobbin. *e-mail: marcello.maresca@astrazeneca.com; jjoung@mgm.harvard.edu

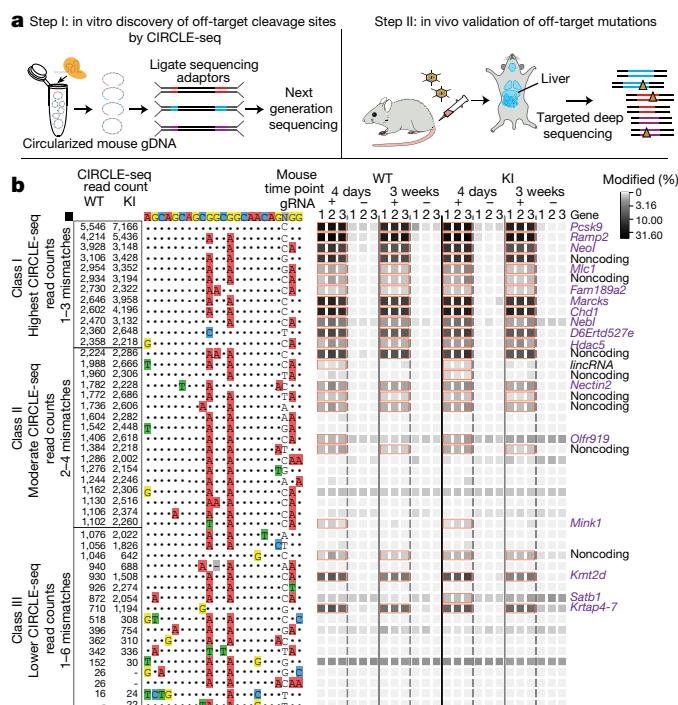


Fig. 1 | Overview and validation of VIVO. a, Schematic illustrating the two-step VIVO method. In step I, CIRCLE-seq identifies off-target sites cleaved in vitro. In step II, the sites identified in step I are assessed in vivo for indel mutations by targeted amplicon sequencing performed with genomic DNA isolated from the livers of nuclease-treated mice. **b**, Assessment of in vivo off-target indels induced by gP-Cas9. Indel frequencies as determined by targeted amplicon sequencing are presented as heat maps for the gP-Cas9 on-target site (black square) and the class I, class II and class III off-target sites (identified from CIRCLE-seq experiments). Each locus was assayed in $n = 3$ biologically independent wild-type (WT) and knock-in (KI) mice (labelled as 1, 2 and 3) using genomic DNA isolated from the liver of mice treated with experimental adenoviral vector that encodes gP-Cas9 (gRNA +) or control adenoviral vector that encodes GFP-Cas9 (gRNA -). Mismatches relative to the on-target site are shown with coloured boxes. CIRCLE-seq read-count numbers for each site are shown. Sites that showed a significant difference between the experimental (gRNA +) and control (gRNA -) samples are outlined with orange boxes and labelled by genomic locus, with coding regions shown in purple text. P values and significance were obtained by fitting a negative binomial generalized linear model (for source data and P values, see Supplementary Table 2).

45 sites we examined, 19 showed significant evidence of indels (mean frequencies ranging from 41.9% to 0.13%) in wild-type and knock-in mouse livers at day four and at week three after infection (Fig. 1b and Supplementary Table 2). Notably, higher CIRCLE-seq read counts generally correlated well with the likelihood of finding indels in vivo: 11 of the 11 class I off-target sites, 5 of the 17 class II sites and 3 of the 17 class III sites contained indels (Fig. 1b). All 19 of these sites had 3 or fewer mismatches relative to the on-target site, with the majority of the sites located within gene-coding sequences (Fig. 1b). Three additional sites—including two sites each containing three mismatches in the spacer and one mismatch in the PAM—showed significant evidence of indel mutations at the four-day time point, but which was no longer significant by three weeks (Supplementary Table 2); the mutation frequencies observed at these sites were around 0.13%, which is close to the limit of detection (0.1%) for next-generation sequencing¹⁰. These data show that Cas9 with a promiscuous gRNA can generate stable, and sometimes high-frequency, off-target mutations in vivo and that VIVO can identify such mutations even with frequencies as low as 0.13%.

Because potential therapeutic applications would not use a promiscuous gRNA with so many closely matched genomic sites, we sought

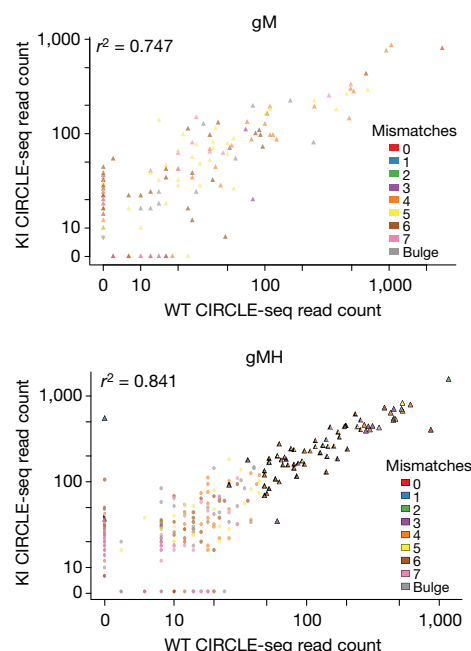


Fig. 2 | Characterization of *Pcsk9*-targeted gRNAs designed to be orthogonal to the mouse genome by CIRCLE-seq. Scatter plots of CIRCLE-seq read counts for off-target cleavage sites identified in vitro with gM-Cas9 and gMH-Cas9 on genomic DNA from wild-type ($n = 1$) and knock-in ($n = 1$) mice (for source data, see Supplementary Tables 3, 4). Each site is colour-coded for the number of mismatches it has relative to the on-target site. Sites represented as triangles were chosen for targeted amplicon sequencing. Correlation r^2 values shown were obtained by linear regression performed using all values in each scatter plot.

to assess the in vivo off-target profiles of Cas9 gRNAs designed to be more orthogonal to the mouse genome. We constructed two additional gRNAs (abbreviated as gM¹¹ and gMH) that are targeted to mouse *Pcsk9* (Extended Data Fig. 6a) but have relatively few closely matched sites in the C57BL6/N mouse genome (Extended Data Table 1). gMH also targets a site (with one mismatch) in the human *PCSK9* open reading frame, which is integrated in the knock-in mouse genome (Extended Data Fig. 6a). Delivery of gM-Cas9 and gMH-Cas9 by adenoviral vectors induced expected genetic alterations in the mouse *Pcsk9* gene and human *PCSK9* transgene as well as corresponding decreases in the levels of *Pcsk9* and *PCSK9* in plasma (Extended Data Fig. 6b, c).

We conducted the first in vitro CIRCLE-seq step of VIVO with gM-Cas9 and gMH-Cas9 on wild-type and knock-in mouse genomic DNA. The on-target mouse *Pcsk9* was identified in all four experiments and the human *PCSK9* transgene site only in the experiment with gMH on knock-in mouse DNA (Fig. 2). We identified fewer off-target sites with gM-Cas9 and gMH-Cas9 (Fig. 2 and Supplementary Tables 3, 4) than we had with gP-Cas9. When using gM, we found 182 off-target sites: 129 with wild-type mouse DNA, 145 with knock-in mouse DNA and 92 sites that were common to both. When using gMH, we found 529 off-target sites: 333 with wild-type mouse DNA, 394 with knock-in mouse DNA and 198 sites that were common to both. All but 2 of the 711 off-target sites that were identified with these gRNAs had 3 or more mismatches, consistent with the higher orthogonality of these gRNAs relative to the mouse genome. For both gRNAs, there were good concordances in CIRCLE-seq read counts between the wild-type and knock-in mice when considering all off-target sites that were identified ($r^2 = 0.747$ and 0.841 for gM and gMH, respectively) (Fig. 2). As when using gP, sites that were found in only one mouse genome generally had low CIRCLE-seq read counts (Fig. 2), and analysis of gM CIRCLE-seq and targeted amplicon sequencing data suggests that these differences are not due to single nucleotide polymorphisms (Extended Data Table 2 and Supplementary Table 5).

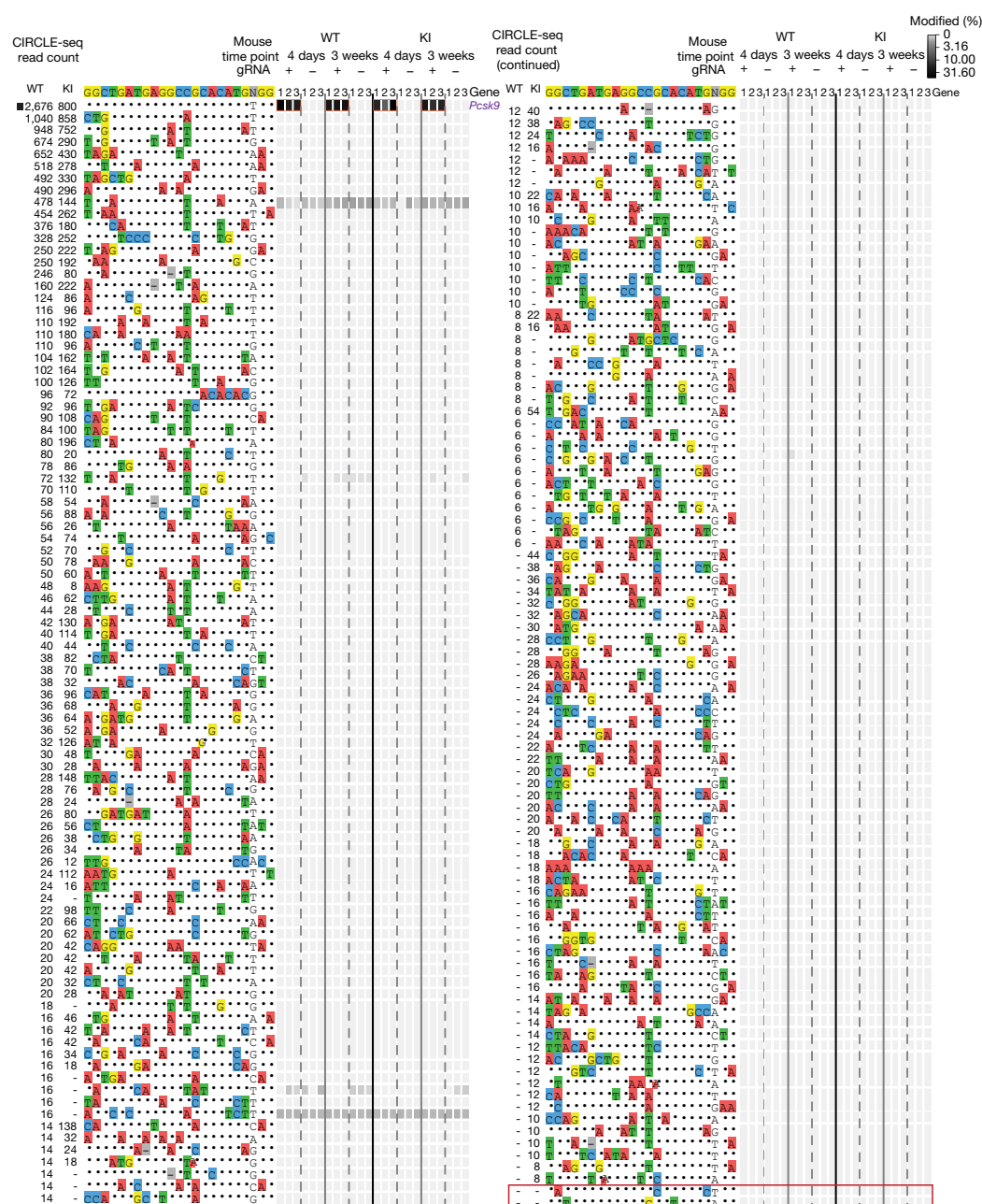


Fig. 3 | Assessment of in vivo off-target indels induced by gM-SpCas9. Indel frequencies determined by targeted amplicon sequencing for the gM-Cas9 on-target site (black square) and 181 off-target sites identified by CIRCLE-seq. Each condition shown was assayed in $n = 3$ biologically independent mice (labelled 1, 2 and 3) using genomic DNA isolated from the liver of mice treated with experimental adenoviral vector that encodes gM-Cas9 (gRNA +) or control adenoviral vector that encodes GFP-Cas9

We conducted the second in vivo step of VIVO for gM and gMH by performing targeted amplicon sequencing of sites found by CIRCLE-seq in liver genomic DNA of adenovirus-treated wild-type and knock-in mice. For gM, we comprehensively examined 181 of the 182 off-target CIRCLE-seq cleavage sites (one site could be amplified but not sequenced; see Methods) and the on-target site from liver DNA collected at day 4 or week 3 after adenovirus infection. Notably, only the gM on-target site showed significant evidence of indels (ranging from 12.6% to 18.5%) (Fig. 3 and Supplementary Table 6); no significant off-target indels were identified at any of the 181 off-target sites in either mouse at either time point. Because CIRCLE-seq identified a large number of potential off-target sites (529 in total) when using gMH, we examined the on-target site and a subset of the off-target sites

(gRNA –). Data are presented as in Fig. 1b. The single site (the on-target site) that was significantly different between the experimental (gRNA +) and control (gRNA –) samples is highlighted with orange boxes. Additional closely matched sites in the mouse genome (not identified from the CIRCLE-seq experiments) examined for indel mutations are boxed in red. For source data and P values (negative binomial), see Supplementary Table 6.

(comprising 69 sites) that had the highest CIRCLE-seq read counts (and up to 6 mismatches). These 69 sites encompassed all but 1 of the CIRCLE-seq sites that had up to 3 mismatches (1 site could be amplified but not sequenced; see Methods) and the human *PCSK9* transgene site (with 1 mismatch) (Fig. 2). This choice was guided by our finding that none of the 19 stable in vivo off-target sites we found for gP-Cas9 had 4 or more mismatches relative to the on-target site (Fig. 1b). Among the 69 sites, we found significant indel mutations at only the on-target mouse gMH site (27.4–43.6% indels) and the human *PCSK9* transgene site bearing one mismatch (20.4–21.7% indels) (Extended Data Fig. 7 and Supplementary Table 7).

To further exclude the possibility that CIRCLE-seq missed any bona fide off-target sites, we also performed targeted amplicon sequencing

of the most closely matched sites in the mouse C57BL/6N genome (containing up to three mismatches in the spacer) that were not identified in our CIRCLE-seq experiments (four sites for gM and ten sites for gMH) (Extended Data Table 1). Three of the gM sites could not be individually selectively amplified and so these were assessed together as a pool (Methods). We did not observe significant indels at any of these sites in all treated mice at both time points (Fig. 3, Extended Data Fig. 7 and Supplementary Tables 6, 7).

To our knowledge, our report provides the first demonstration that CRISPR–Cas nucleases can robustly induce off-target mutations in vivo. Previous in vivo studies have reported no or very few off-target mutations, but used the cell-based ‘genome-wide unbiased identification of double-strand breaks enabled by sequencing’ (GUIDE-seq) method^{12–14} or other in silico approaches that have not been validated to effectively identify these sites in vivo (see Supplementary Discussion). By contrast, VIVO enabled the robust and sensitive identification of off-target sites in vivo, even those with mutation frequencies as low as about 0.13%. The high sensitivity of CIRCLE-seq is most probably what enabled the identification of a superset of all potential off-target sites, including those actually mutated in vivo (Supplementary Discussion). The detection limit of VIVO—as with all existing methods—is bounded by the current error rate of next-generation sequencing for indels (approximately 0.1%). In addition, we did not attempt to detect large-scale chromosomal rearrangements (translocations, inversions or large deletions) but we expect the frequencies of such alterations to be no higher than that of any off-target indels. Other methods¹⁵ might be used in future studies to test for these alterations.

VIVO defines a pathway for assessing the in vivo genome-wide specificities of CRISPR–Cas nucleases. Our work suggests that gRNAs should be designed to have the lowest possible number of closely matched genomic sites (with three or fewer mismatches), which can be done using existing in silico tools¹⁶. Such gRNAs can be assessed in step 1 of VIVO to identify those that have a reasonable number (<100–200) of in vitro off-target cleavage sites and then these sites can be comprehensively examined for indels in vivo using targeted amplicon sequencing in step 2 of VIVO. We also recommend examination of any other closely matched genomic sites (fewer than four mismatches) that are not identified by CIRCLE-seq. Persistent off-target mutations might be further reduced using various strategies for improving the genome-wide specificities of CRISPR–Cas nucleases (Supplementary Discussion).

We believe VIVO sets an important standard for defining in vivo off-target effects of gene-editing nucleases. The approach should be generalizable to non-CRISPR gene-editing nucleases, to non-mammalian organisms and to other delivery methods (Supplementary Discussion). VIVO should also be useful for characterizing the in vivo specificities of engineered CRISPR–Cas9 variants and other CRISPR–Cas orthologues (Supplementary Discussion) and for assessing the effectiveness of methods for reducing the off-target effects of these nucleases. We expect our findings will advance research and methods that will further spur the clinical translation of in vivo genome-editing therapeutic strategies.

Note added in proof: a recent study¹⁷ that investigated large-scale chromosome alterations resulting from CRISPR–Cas9 editing found that the frequencies of these alterations were indeed no higher than those of the off-target indels we report here.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0500-9>.

Received: 27 February 2018; Accepted: 23 July 2018;

Published online 12 September 2018.

- Musunuru, K. The hope and hype of CRISPR–Cas9 genome editing: a review. *JAMA Cardiol.* **2**, 914–919 (2017).
- Fellmann, C., Gowen, B. G., Lin, P. C., Doudna, J. A. & Corn, J. E. Cornerstones of CRISPR–Cas in drug discovery and therapy. *Nat. Rev. Drug Discov.* **16**, 89–100 (2017).

- Komor, A. C., Badran, A. H. & Liu, D. R. CRISPR-based technologies for the manipulation of eukaryotic genomes. *Cell* **168**, 20–36 (2017).
- Koo, T. & Kim, J. S. Therapeutic applications of CRISPR RNA-guided genome editing. *Brief. Funct. Genomics* **16**, 38–45 (2017).
- Cornu, T. I., Mussolino, C. & Cathomen, T. Refining strategies to translate genome editing to the clinic. *Nat. Med.* **23**, 415–423 (2017).
- Dunbar, C. E. et al. Gene therapy comes of age. *Science* **359**, eaan4672 (2018).
- Tsai, S. Q. & Joung, J. K. Defining and improving the genome-wide specificities of CRISPR–Cas9 nucleases. *Nat. Rev. Genet.* **17**, 300–312 (2016).
- Tsai, S. Q. et al. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).
- Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
- Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
- Ding, Q. et al. Permanent alteration of PCSK9 with in vivo CRISPR–Cas9 genome editing. *Circ. Res.* **115**, 488–492 (2014).
- Yin, H. et al. Structure-guided chemical modification of guide RNA enables potent non-viral in vivo genome editing. *Nat. Biotechnol.* **35**, 1179–1187 (2017).
- Yin, H. et al. Therapeutic genome editing by combined viral and non-viral delivery of CRISPR system components in vivo. *Nat. Biotechnol.* **34**, 328–333 (2016).
- Gao, X. et al. Treatment of autosomal dominant hearing loss by in vivo delivery of genome editing agents. *Nature* **553**, 217–221 (2018).
- Giannoukos, G. et al. UDI-TASTM, a genome editing detection method for indels and genome rearrangements. *BMC Genomics* **19**, 212 (2018).
- Bae, S., Park, J. & Kim, J. S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).
- Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* **36**, 765–771 (2018).

Acknowledgements J.K.J. is supported by the Desmond and Ann Heathwood MGH Research Scholar Award. J.K.J., M.L.B. and J.A.G. were supported by a sponsored research agreement with AstraZeneca. L.P. is supported by a National Human Genome Research Institute (NHGRI) Career Development Award (R00HG008399). J.K.J., M.J.A. and J.M.-L. are supported by a National Institutes of Health Maximizing Investigators' Research Award (MIRA) (R35 GM118158). J.K.J., L.P. and K.C. are supported by the Defense Advanced Research Projects Agency (HR0011-17-2-0042). We thank M. Snowden, S. Platz and S. Rees for resource allocation from AstraZeneca Research Funds. We thank J. Y. Hsu for discussions and input.

Reviewer information Nature thanks F. Urnov and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions P.A., M.J.P., T.B. and M.B. executed intravenous tail vein injections. P.A. and T.B. coordinated vena saphena blood sampling and performed plasma extractions. P.A., M.J.P., A.C., T.B., M.B., M.M. and R.N. performed in vivo terminations and organ collection. P.A. performed Surveyor assays for genomic DNAs of mouse livers. P.A. and T.B. performed ELISA for Pcsk9 protein detection in plasma samples. M.L.B., J.A.G., S.Q.T. and N.T.N. performed the CIRCLE-seq and targeted amplicon sequencing experiments. J.M.-L., K.C., S.P.G., L.P. and M.J.A. performed bioinformatic and computational analysis of the off-target experiments. M.A.F. generated AstraZeneca proprietary software for gRNA identification. P.A. designed gRNAs with help of M.A.F., and validated their functional activity. F.S. performed mouse phenotypic characterization. P.A., M.L.B., S.Q.T., M.M., M.B.-Y., A.C., R.N., M.D.F., L.M.M. and J.K.J. conceived of and designed the study. P.A., M.L.B., M.M. and J.K.J. organized and supervised experiments. P.A., M.L.B., J.A.G., M.M. and J.K.J. prepared the manuscript with input from all authors.

Competing interests J.K.J. has financial interests in Beam Therapeutics, Blink Therapeutics, Editas Medicine, Endcadia, Monitor Biotechnologies (formerly known as Beacon Genomics), Pairwise Plants, Poseida Therapeutics and Transposagen Biopharmaceuticals. M.J.A. and S.Q.T. have financial interests in Monitor Biotechnologies. J.K.J.'s and M.J.A.'s interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies. S.Q.T. and J.K.J. are co-inventors on a patent describing the CIRCLE-seq method. P.A., M.D.F., M.J.P., M.A.F., F.S., M.B., R.N., M.B.Y. and M.M. are employees and shareholders of AstraZeneca. L.M.M. is an employee and shareholder of GE Healthcare and a shareholder of AstraZeneca.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0500-9>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0500-9>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to M.M. or J.K.J.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

gRNA design. gP was identified by searching for an on-target sequence within mouse *Pcsk9* (ENSMUSG000044254) exons one to three that showed a high number of closely matched sites (two or fewer mismatches to the on-target site) in the mouse genome. gMH was designed by searching for a gRNA that can cleave both mouse *Pcsk9* and human *PCSK9*, and that showed a perfect alignment to mouse *Pcsk9* and up to two nucleotides mismatch to human *PCSK9* (ENSG00000169174) at least eight nucleotides distal from the PAM. For both gRNA designs, AstraZeneca proprietary software was used as an in silico tool, which was developed based on the codebase of the Wellcome Trust Sanger Institute (WGE: <http://www.sanger.ac.uk/htgt/wge/>)¹⁸ with the addition of the NAG PAM motif and NGG in the alignments for potential off-targets⁹. GRCm38/mm10 and GRCh38/hg38 genomes were used as reference for alignments. The gM targeted to mouse *Pcsk9* has previously been described¹¹.

Adenoviral constructs. Adenoviruses that express Cas9 and gRNAs (Ad-Cas9-gM, Ad-Cas9-gMH and Ad-Cas9-gP) were generated by Vector Biolabs (Malvern). Cas9 and gRNAs were expressed from chicken β -actin hybrid (CBh) and U6 promoters, respectively, in a replication-deficient adenoviral-serotype 5 (dE1/E3) backbone. A negative control adenovirus (Ad-Cas9-GFP) that expresses Cas9 and GFP from the CBh and CMV promoters, respectively, but no gRNA was also generated.

Animal studies. All animal experiments were approved by the AstraZeneca internal committee for animal studies as well as the Gothenburg Ethics Committee for Experimental Animals (license number: 162-2015+), compliant with EU directives on the protection of animals used for scientific purposes.

C57BL/6N mice (Charles River) were individually housed in a temperature ($21 \pm 2^\circ\text{C}$) and humidity ($55 \pm 15\%$) controlled room with a 12:12-h light:dark cycle. R3 diet (Lactamin AB) and tap water were provided ad libitum. Cage bedding and enrichments include spen chips, shredded paper, gnaw sticks and a plastic house.

A humanized hypercholesterolaemia mouse model was generated by liver-specific overexpression of human *PCSK9* in C57BL/6N mice (A.C. et al., manuscript submitted). In brief, the knock-in mouse was generated by cloning the human *PCSK9* open reading frame downstream of the mouse albumin promoter from albumin-Cre mice, in a vector designed to target the mouse *Rosa26*. Founder C57BL/6N hPCSK9 heterozygous males were crossed with C57BL/6N females to generate experimental animals, which are littermates with two genotypes: C57BL/6N hPCSK9KI^{+/−} (referred to as knock-in) and wild-type C57BL/6N hPCSK9KI^{−/−} (referred to as wild-type) (Extended Data Fig. 3).

For in vivo *Pcsk9* gene editing, nine- to eleven-week-old male mice received a tail vein injection with a dose of 1×10^9 infection units (IFU) of adenovirus (Ad-Cas9-gM, Ad-Cas9-gMH, Ad-Cas9-gP or Ad-Cas9-GFP) in 200 μl diluted with phosphate-buffered saline. Peripheral blood was sampled before virus administration (baseline), a week after virus administration and at termination (four days or three weeks after virus administration). Animals were euthanized by cardiac puncture under isoflurane anaesthesia at the experimental endpoint. The organs—including liver, spleen, lungs, kidney, muscle, brain and testes—were dissected, snap-frozen in liquid nitrogen and stored at -80°C until further analyses.

Ten milligrams of frozen liver tissue was lysed to obtain 30–50 μg genomic DNA using the Gentra Puregene Tissue kit (Qiagen). In vivo gene-editing efficiency was evaluated using Surveyor mismatch cleavage assay (Integrated DNA Technologies, BVBA) (using primers: for gP, GAGGCCGAAACCTGATCCTT and CTTAGAGACCACAGACGGC; for gM, GGAGGACACGTTTCTGCAT and CTGCTGCTGTTGCTGCTAC; for gMH mouse locus, GACTTTGTGA AGGCTGGGA and TGCATGGAGCAATGCAGAGA; for gMH human transgene, TAGCCTTGCGTTCGAGGAG and CATTCTCGAAGTCGGTGACCA; with amplicon sizes 619, 415, 349, 495 base pairs, respectively) and targeted deep sequencing (primers listed in Supplementary Tables 2, 6 and 7).

Animals were randomized based on their weights measured before the experiments. The investigators were blinded to group allocation during data collection and analysis. No sample size calculation was performed. Sample size was determined to generate triple independent samples for comparisons between groups sufficient to perform statistical tests.

Assessment of human *PCSK9* and/or mouse *Pcsk9* protein levels in plasma. Peripheral blood was collected in EDTA-coated capillary tubes from *vena saphena* during the course of the study and by cardiac puncture at the time of termination. Samples were kept on ice for up to 2 h before extraction of plasma by centrifugation at 10,000 rpm for 20 min at 4°C . Plasma was stored at -80°C until the samples were analysed. Levels of human *PCSK9* and mouse *Pcsk9* in plasma were determined with a standard ELISA kit (DPC900 and MPC900; R&D Systems) according to the manufacturer's instructions. Prior to the assay, plasma samples were diluted 1:800 and 1:1,000 for human *PCSK9* and mouse *Pcsk9*, respectively.

Reference genome for CIRCLE-seq, CRISPResso and Cas-OFFinder. Build 38 of the C57BL/6N genome, sequenced by the Sanger Mouse Genomes Project

(<http://csbio.unc.edu/CCstatus/index.py?run=PseudoOld>), was used as the reference genome for the experiments of this report. The human *PCSK9* gene DNA sequence was inserted into the mouse genome as an extra chromosome in the reference and was named as 'chrPCSK9KI'.

CIRCLE-seq. CIRCLE-seq was performed experimentally as previously described⁸. Data were processed using v.1.1 of the CIRCLE-Seq analysis pipeline⁸ (<https://github.com/tsailabSJ/circleseq>) with parameters: 'window_size: 3; mapq_threshold: 50; start_threshold: 1; gap_threshold: 3; mismatch_threshold: 7; merged_analysis: False; variant_analysis: True'. The gP off-target sites were adjusted for mapping artefacts in highly repetitive regions by consolidating contiguous sites whose mapping positions differed by 5 base pairs or less. The off-target sequences reported for these consolidated sites were obtained by performing the local alignment used in the CIRCLE-seq pipeline.

Targeted amplicon deep sequencing of off-targets. Genomic DNA from liver tissue of adenovirus-injected mice was extracted at day 4 and at week 3 post-treatment for indel analysis. As detailed in the text, we validated off-targets identified by CIRCLE-seq by selecting sites with read counts above 50% of the on-target and a variety of lower-ranked sites (containing up to 6 mismatches relative to the on-target) for targeted deep sequencing. In addition, we ruled out the possibility that CIRCLE-seq was missing potential off-target sites identified using in silico tools by sequencing all sites containing up to 3 mismatches, identified by Cas-OFFinder¹⁶ for gM and gMH. All sites we analysed were amplified from 150 ng of input genomic DNA (approximately 5×10^4 genomes) with Phusion Hot Start Flex DNA polymerase (New England Biolabs). PCR products were purified using magnetic beads made as previously described¹⁹, quantified using a QuantiFluor dsDNA System kit (Promega), normalized to 10 ng/ μl per amplicon and pooled. Pooled samples were end-repaired and A-tailed using an End Prep Enzyme Mix and reaction buffer from NEBNext Ultra II DNA Library Prep Kit for Illumina, and ligated to Illumina TruSeq adapters using a ligation master mix and ligation enhancer from the same kit. Library samples were then purified with magnetic beads made as previously described¹⁹, size-selected using PEG/NaCl SPRI solution (KAPA Biosystems), quantified using droplet digital PCR (BioRad) and loaded onto an Illumina MiSeq for deep sequencing. To analyse amplicon sequencing of potential on- and off-targets, we used CRISPResso software²⁰ v.1.0.11 (<https://github.com/lucapinello/crispresso>) with the following parameters: '-q 30 --ignore_substitutions --hide_mutations_outside_window_NHEJ'.

For each of the 45 gP off-target sites we examined, we obtained 10,000 or more sequencing reads in at least 2 samples for treated and control samples at all time points (Supplementary Table 2). One potential gM off-target site (chr15:98037617-98037640) and one potential gMH off-target site (chr15:4878177-4878200) were amplified but could not be successfully sequenced. The problematic gM site was amplified with two different sets of primers and both amplicons failed to sequence. The gMH site is in a highly repetitive area with low complexity and we were unable to differentiate this site from other sites in the genome; therefore, the site was removed from analysis. For all of the gM off-target sites we were able to sequence, we obtained 10,000 or more sequencing reads in at least 2 samples for treated and control samples at all time points (Supplementary Table 6). For all but 1 of the gMH off-target sites we were able to sequence, we obtained 10,000 or more sequencing reads in at least 2 samples for treated and control samples at all time points (Supplementary Table 7). One of the gMH off-target sites we sequenced (chr17:33501685-33501708) did not reach the 10,000 read threshold for any samples or time points but read counts ranged from 2,509 to 9,149. We were unable to individually and selectively amplify three sites that were identified in silico as being highly similar to the gM on-target site but that were not identified by CIRCLE-seq, owing to their sequence similarities: these sites were chr14:25878231-25878254, chr14:26018001-26018024 and chr14:26157615-26157638. Therefore, for these three sites the read counts were pooled into one amplicon that encompasses all locations and that is labelled as 'chr14:pooled' in Supplementary Table 6.

Targeted amplicon deep sequencing of wild-type and knock-in untreated mice. Genomic DNA from liver tissue of untreated mice—the same mice upon which CIRCLE-seq was performed—was amplified at sites that contained no reads in either the knock-in or wild-type CIRCLE-seq for gM and deep sequenced to look for single nucleotide polymorphisms (Supplementary Table 5). Primers used are the same as in Supplementary Table 6. Sites were analysed with CRISPRessoPooled with the following parameters: '--cleavage_offset -7 --window_around_sgrna 13', to perform a focused variant analysis in the spacer.

Cas-OFFinder. Identification of potential off-targets by Cas-OFFinder¹⁶ (<https://github.com/snugeli/cas-offinder>, v. 2.4) was done using the off-line version, allowing up to 7 mismatches and non-canonical PAMs. We then restricted the output to the sites with at most 6 mismatches in the spacer and at most 1 mismatch in the PAM.

Non-reference genetic variation. samtools (mpileup and bcftools, v.1.3.1²¹) was used to discover non-reference genetic variation at the off-target sites identified by CIRCLE-seq. Positions with a genotype-quality score greater than 5 and depth

of at least 3 were considered as potential variants if they did not fall adjacent to the cleavage site or at the edge of the reads, and were not located in a highly repetitive region with poor mapping quality.

Statistical analysis of levels of protein in plasma. Data visualization and statistical analyses for plasma protein measurements were performed using GraphPad Prism 7.02. Protein levels after the adenoviral administration were normalized to baseline levels, and values for gRNA treatment groups were compared with the control treatment group. Comparisons between groups were performed using two-way ANOVA test followed by Sidak's or Dunnett's two-sided adjusted multiple comparisons test, depending on the number of comparison groups. $P < 0.05$ was considered to be statistically significant. The level of significance in all graphs is represented as follows: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$. Exact P values, confidence intervals and effect size are presented in the Source Data.

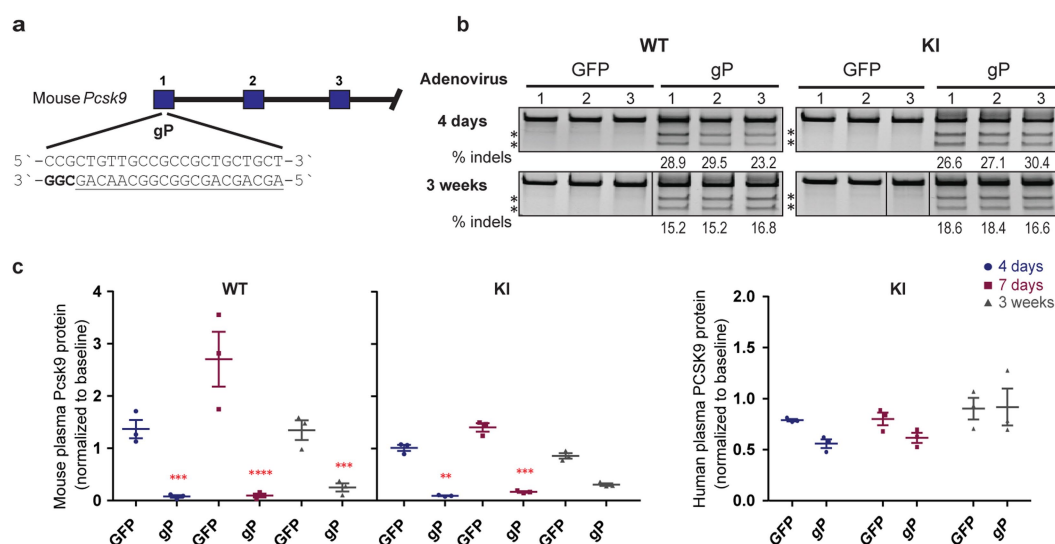
Statistical analysis of targeted amplicon deep-sequencing data. P values were obtained by fitting a negative binomial generalized linear model (function MASS:glm.nb in R version 3.4.2 with parameter $\text{init.theta} = 1$ and with the logarithm of the total number of reads as the offset) to the control and nuclease-treated samples for each evaluated site with at least one non-zero indel count among the nuclease-treated samples. To avoid convergence issues, we added 1 to all the indel counts and confirmed that rerunning the models without the addition of the 1 did not result in any additional significant off-target sites. We adjusted for multiple

comparisons using the Benjamini and Hochberg method (function p.adjust in R version 3.4.2). Multiple testing adjustment was performed within strata defined by gRNA, mouse background and time point. We considered the indel percentage in the gRNA-Cas9-treated replicates to be significantly greater than the indel percentage in the GFP-Cas9-treated controls if the adjusted P value was less than 0.1, the nuclease-treatment coefficient was greater than zero and the median indel frequency of the treated replicates was greater than 0.1%.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

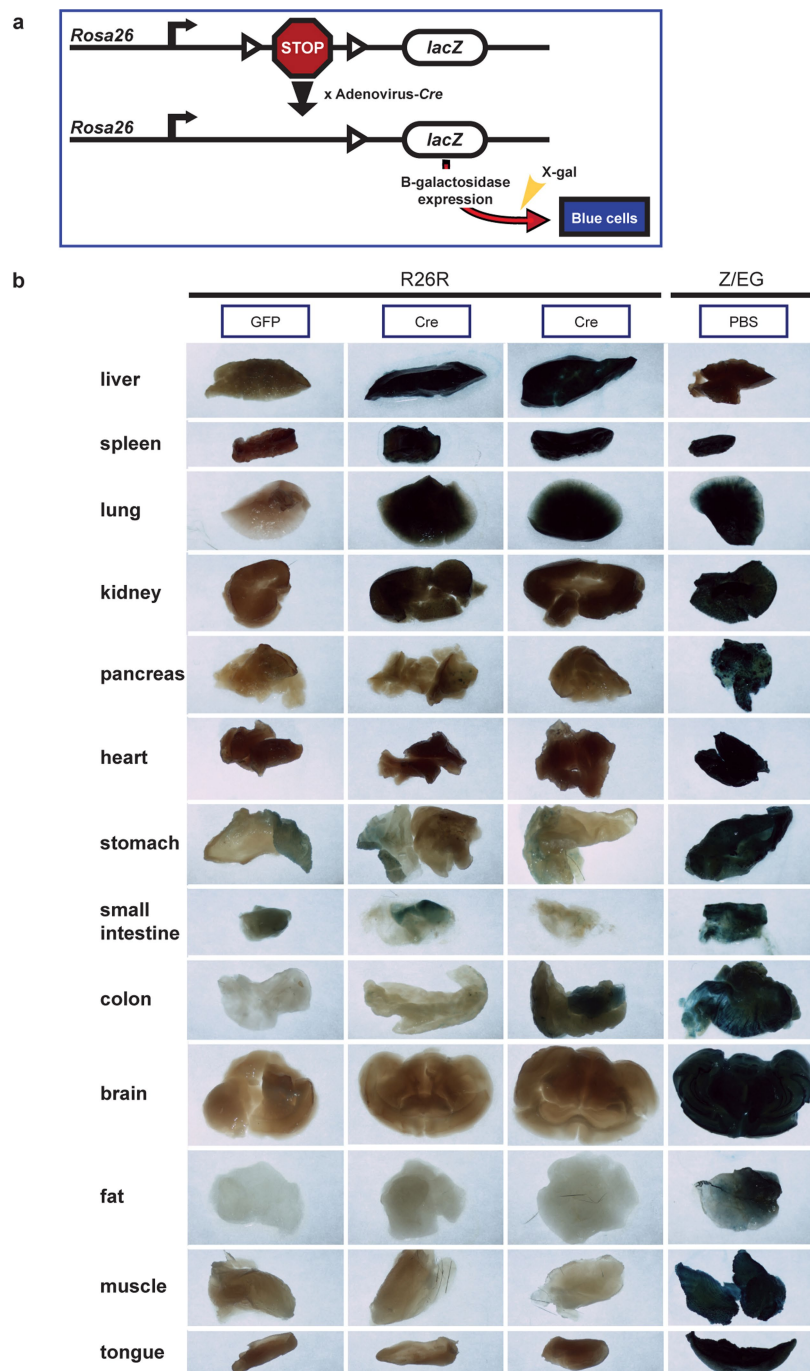
Data availability. Sequence data that support the findings of this study have been deposited with SRA accession number SRP151131. All other data that support the findings of this study are available from the corresponding authors upon reasonable request.

18. Hodgkins, A. et al. WGE: a CRISPR database for genome engineering. *Bioinformatics* **31**, 3078–3080 (2015).
19. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
20. Pinello, L. et al. Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat. Biotechnol.* **34**, 695–697 (2016).
21. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).



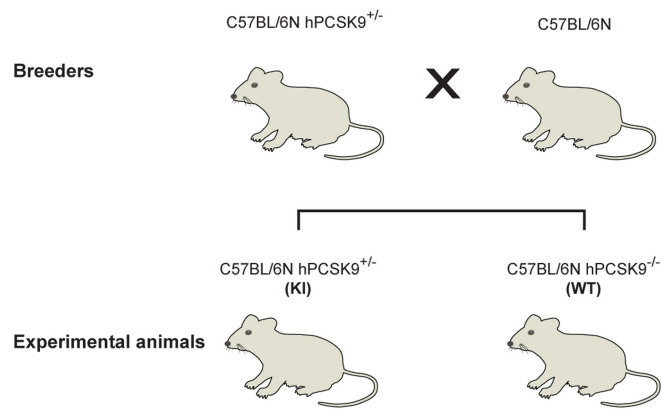
Extended Data Fig. 1 | gP-Cas9 efficiently mutates the mouse *Pcsk9* gene and reduces levels of *Pcsk9* protein in plasma in vivo. **a**, gP was designed to target a sequence within exon 1 of the mouse *Pcsk9* gene that has many closely related genomic sites (that is, those with 1–3 mismatches relative to the on-target site; Extended Data Table 1). Blue bars indicate exons for the mouse genomic region. **b**, Surveyor assay and next-generation DNA sequencing data demonstrate efficient in vivo modification of the on-target mouse *Pcsk9* gene site in mouse liver by gP-Cas9. Assays were performed on day 4 and on week 3 after the administration of adenoviral vectors that encode gP-Cas9 (gP) or negative control GFP-Cas9 (GFP). For each time point, the assays used genomic DNA isolated from livers of $n = 3$ biologically independent wild-type C57BL/6N (WT) mice or C57BL/6N-derived mice containing a single copy of the human *PCSK9* open reading frame under albumin promoter, knocked into the *Rosa26* locus (KI). Asterisks indicate the cleaved PCR products expected after treatment with Surveyor nuclease. Percentages show the frequencies of indel mutations determined by targeted amplicon sequencing using next-generation sequencing; these

are the same values shown for the on-target site in Fig. 1b. Lines divide lanes taken from different locations on the same gel. For source data for Surveyor assays and targeted amplicon sequencing, see Supplementary Fig. 1 and Supplementary Table 2, respectively. **c**, Mouse *Pcsk9* protein levels in plasma measured in $n = 3$ biologically independent wild-type and knock-in mice and human PCSK9 protein levels in plasma measured in $n = 3$ biologically independent knock-in mice, after nuclease treatment. Protein levels were assessed on day 4, 7 and week 3 after administration of gP or control GFP adenoviral vectors and normalized to baseline levels. Significant differences between experimental and control groups were determined using two-way ANOVA and Sidak's two-sided adjusted multiple comparisons test; $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$. See Source Data for Extended Data Fig. 4b for exact adjusted P values. All values are presented as group means, and error bars represent standard error of the mean. The enhanced reduction of levels of *Pcsk9* in plasma relative to the frequency of observed *Pcsk9* genetic alteration is consistent with previously published studies (Supplementary Discussion).

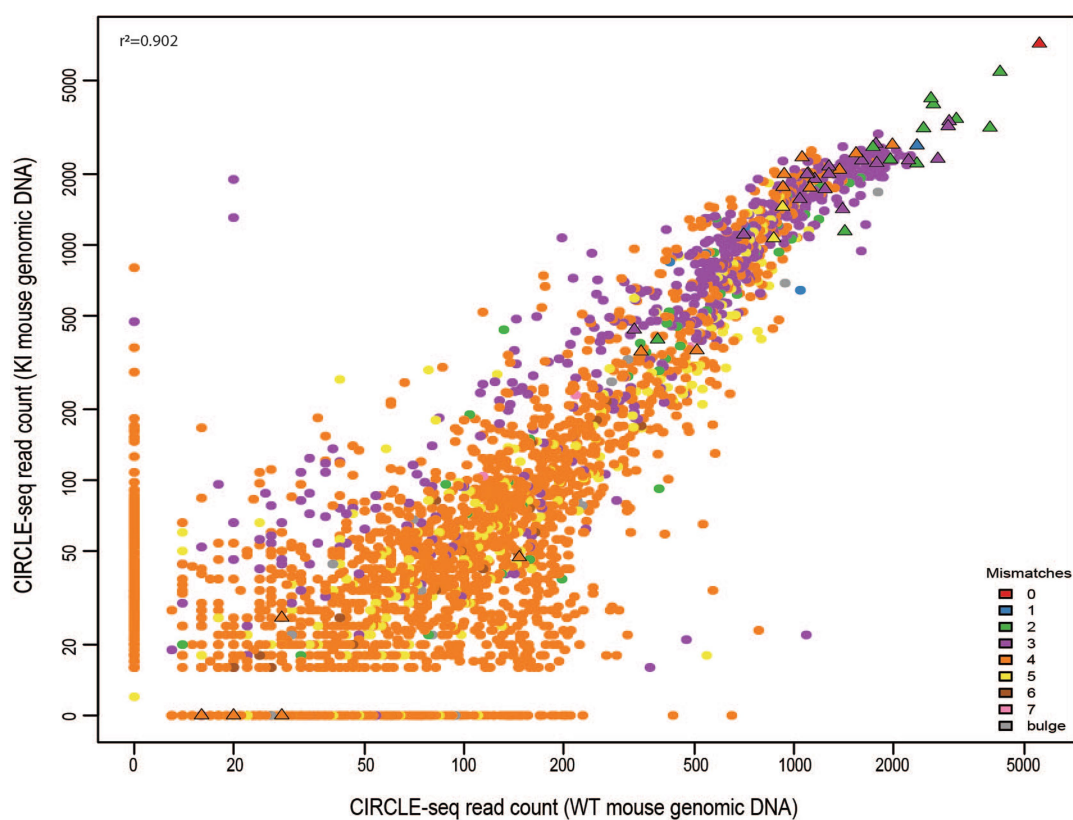


Extended Data Fig. 2 | Bio-distribution studies of adenovirus-serotype 5 in mice. a, Schematic of integrated reporter construct in R26R mice used to assess delivery of Cre recombinase using adenovirus serotype 5 vector. Cre-mediated excision of a loxP-flanked transcriptional stop signal upstream of a *lacZ* gene results in expression of β -galactosidase enzyme. β -Galactosidase expression can be quantified by staining dissected tissues with X-gal, a compound that turns blue when cleaved by this enzyme. **b**, Quantification of β -galactosidase expression in sections of various dissected organs from $n = 2$ biologically independent R26R mice

intravenously injected with adenovirus serotype 5 vector encoding Cre. Matched organs sections from a R26R mouse intravenously injected with an adenovirus serotype 5 vector encoding GFP were used to determine background staining levels and serve as a negative control. Matched organ sections from Z/EG mice that constitutively express *lacZ* (β -galactosidase) and intravenously injected with PBS (rather than adenovirus) were used to provide positive staining controls. All mice were evaluated one week after adenovirus or PBS injection. The experiment was performed once.

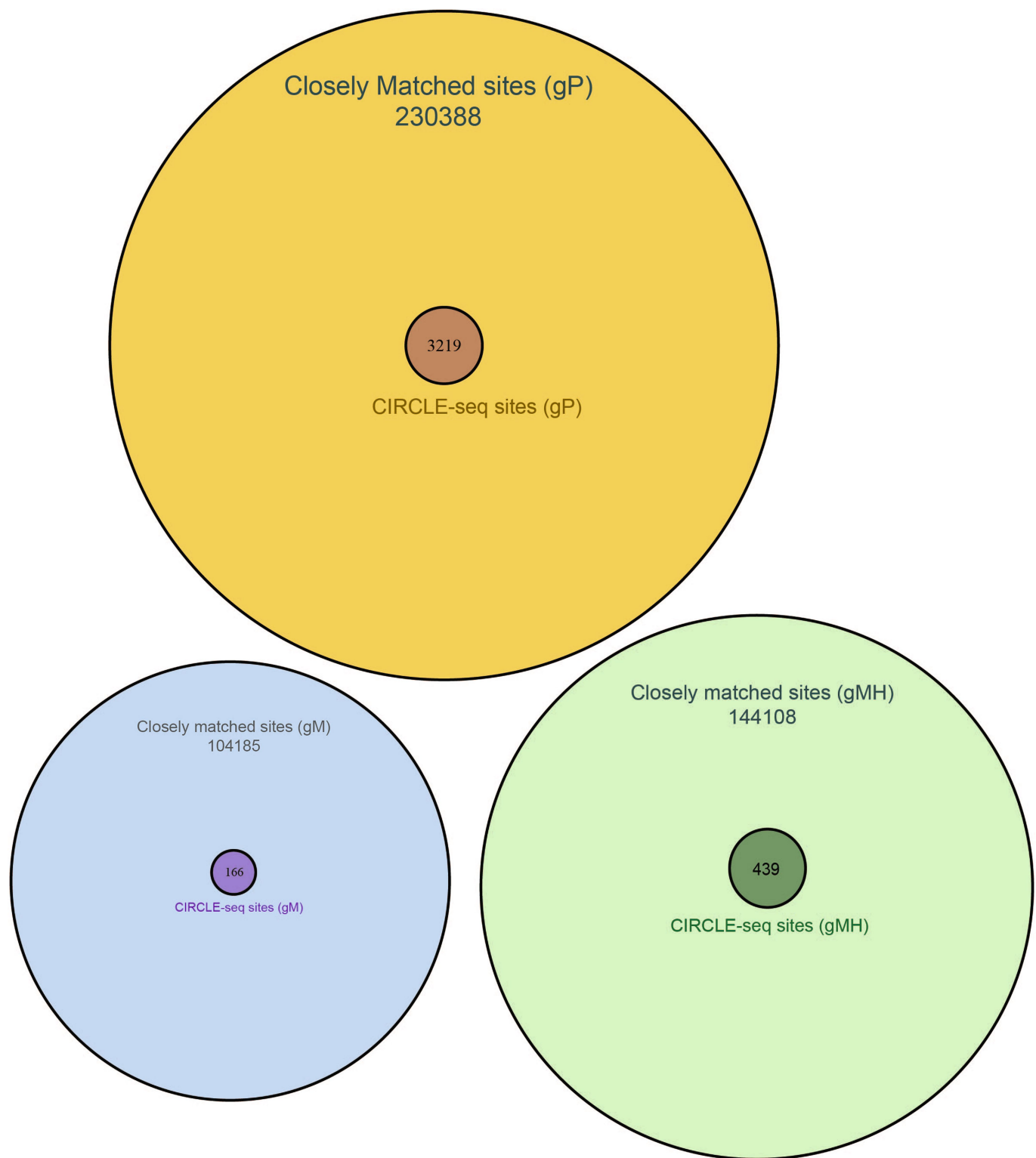


Extended Data Fig. 3 | Breeding strategy for generating experimental mice containing human *PCSK9* open reading frame knocked into the *Rosa26* locus. C57BL/6N-derived mouse line containing a single copy of the human *PCSK9* open reading frame knocked into the *Rosa26* locus (C57BL/6N hPCSK9KI^{+/-}) are used for breeding with C57BL/6N mice. Offspring yielded experimental animals that are C57BL/6N hPCSK9KI^{+/-} (referred to as knock-in) and C57BL/6N hPCSK9KI^{-/-} (referred to as wild type) males.



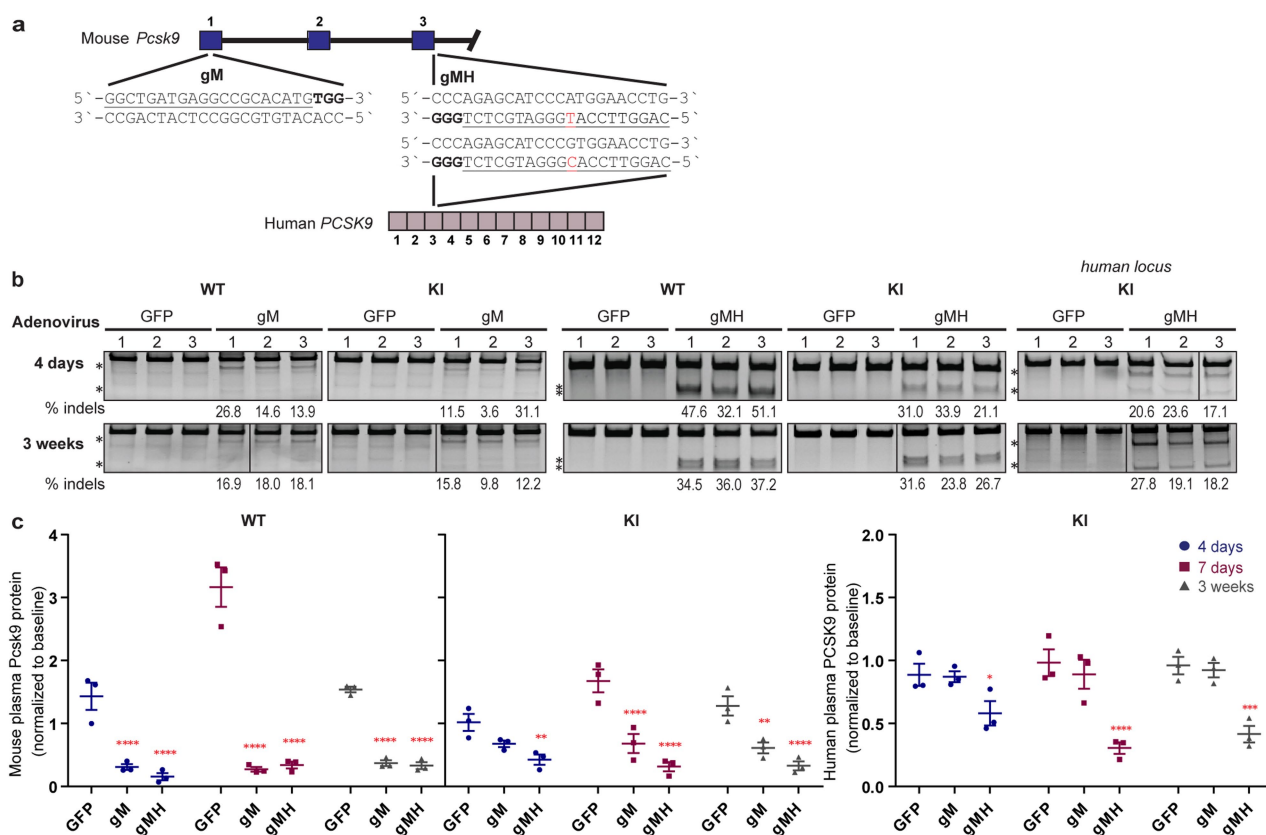
Extended Data Fig. 4 | Scatter plot of CIRCLE-seq read counts for sites identified with gP-Cas9 on genomic DNA from $n = 1$ wild-type and $n = 1$ knock-in mice. Read counts are shown on a logarithmic scale and colours indicate the number of mismatches in each off-target site relative

to the on-target site. Sites shown as triangles were chosen for targeted amplicon sequencing. The correlation r^2 value obtained using all values in the scatter plot is shown in the upper left-hand corner and was obtained using a linear regression.



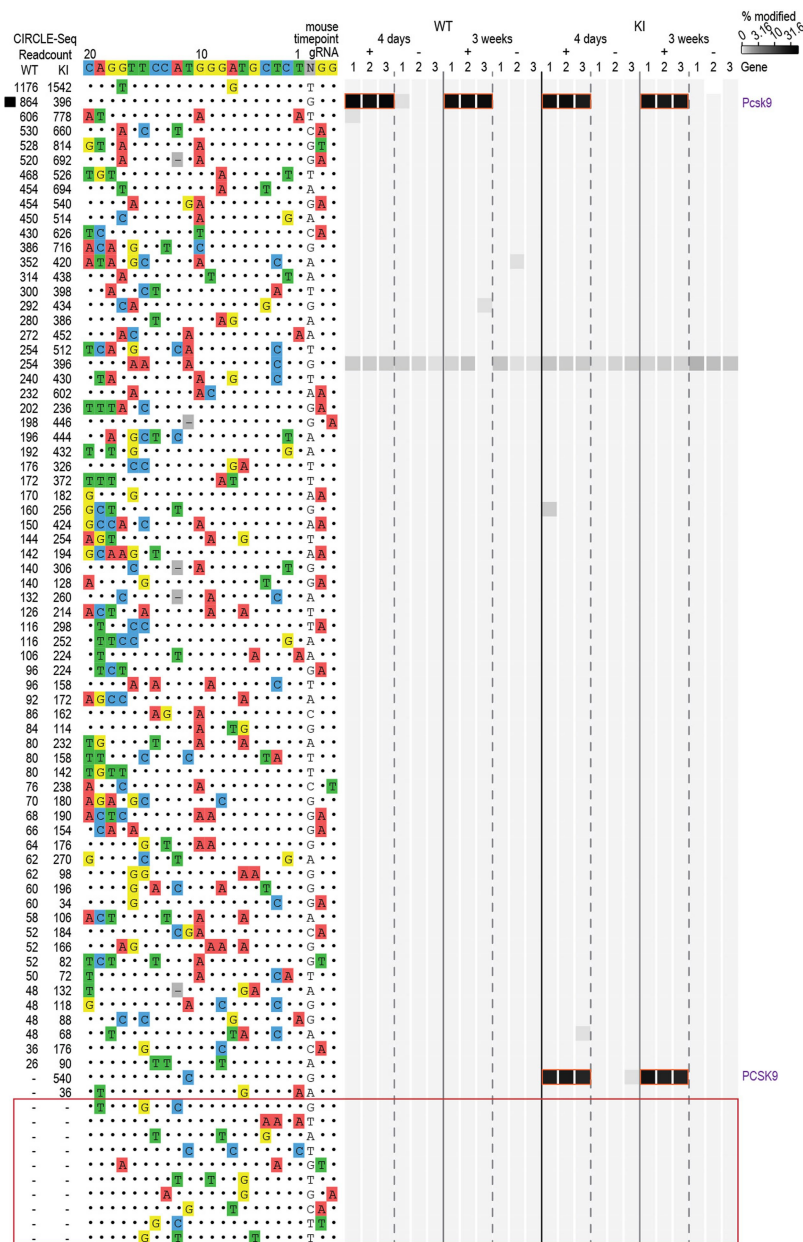
Extended Data Fig. 5 | Comparison of closely matched sites identified in silico and off-target cleavage sites identified by CIRCLE-seq. Venn diagrams comparing off-target cleavage sites in mouse genomic DNA identified by CIRCLE-seq experiments with closely matched sites (up

to six mismatches relative to the on-target site) in the mouse genome identified in silico by Cas-OFFinder are shown for the Cas9 gRNAs gP, gM and gMH.



Extended Data Fig. 6 | Genetic and phenotypic alterations induced by delivery of gM-Cas9 and gMH-Cas9 in vivo. **a**, Sequence and location of the Cas9 gM (mouse) and gMH (mouse and human) target sites in the endogenous mouse *Pcsk9* gene and human *PCSK9* transgene inserted at the mouse *Rosa26* locus. The single base position that differs between the gMH target sites in the mouse *Pcsk9* gene and the human *PCSK9* transgene is highlighted in red. Blue bars indicate exons for the mouse genomic region and purple bars represent exons for the human genomic locus; the PAM sequence for the sites is in bold and the spacer sequence is underlined. **b**, Surveyor assay and next-generation DNA sequencing data demonstrate efficient in vivo modification of the on-target endogenous mouse *Pcsk9* site and human *PCSK9* transgene in mouse liver. Assays were performed at day 4 and at week 3 after administration of adenoviral vectors that encode gM and Cas9 (gM), gMH and Cas9 (gMH) or GFP and Cas9 (GFP) using genomic DNA isolated from livers of $n = 3$ biologically independent wild-type and knock-in mice. Asterisks indicate the cleaved PCR products expected following treatment with Surveyor nuclease. Percentages show the frequencies of indel mutations determined by

targeted amplicon sequencing using next-generation sequencing; these are the same values shown for the on-target sites in Fig. 3 and Extended Data Fig. 7. Lines divide lanes taken from different locations on the same gel. For source data for Surveyor assays, see Supplementary Fig. 1. For source data for targeted amplicon sequencing, see Supplementary Tables 6 and 7 for gM and gMH, respectively. **b**, Mouse *Pcsk9* protein levels measured in plasma in $n = 3$ biologically independent wild-type and knock-in mice, and human *PCSK9* protein levels measured in plasma in $n = 3$ biologically independent knock-in mice after CRISPR-Cas nuclease treatment. Protein levels in plasma were assessed at day 4, 7 and week 3 after the administration of gM, gMH or control GFP adenoviral vectors and normalized to baseline levels at each time point. Significant differences between groups were determined using two-way ANOVA and Dunnett's two-sided adjusted multiple comparisons test; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$. See Source Data for exact adjusted P values. Values are presented as group means, error bars represent standard errors of the mean.



Extended Data Fig. 7 | Assessment of in vivo off-target indel mutations induced by gMH-Cas9. Indel mutation frequencies determined by targeted amplicon sequencing (using high-throughput sequencing) are presented as heat maps for the gMH-Cas9 on-target site (black square) and 63 off-target sites identified from CIRCLE-seq experiments. Each locus was assayed in $n = 3$ biologically independent mice (labelled 1, 2 and 3) using genomic DNA isolated from the liver of wild-type and knock-in mice treated with experimental adenoviral vector that encodes gMH-Cas9 (gRNA +) or control adenoviral vector GFP-Cas9 (gRNA -). For each site, mismatches relative to the on-target site are shown with coloured boxes and bases in the spacer sequence and are numbered from 1 (most proximal to the PAM) to 20 (most distal from the PAM). The number

of read counts found for each site from the CIRCLE-seq experiments on wild-type and knock-in mouse genomic DNA are shown in the left columns (ranked from highest to lowest based on counts in the wild-type genomic DNA CIRCLE-seq experiment). Each box in the heat map represents a single sequencing experiment. Sites that were significantly different between the experimental (gRNA +) and control (gRNA -) samples are highlighted with an orange outline around the boxes. Additional closely matched sites in the mouse genome (not identified from the CIRCLE-seq experiments) that were examined for indel mutations are boxed in red at the bottom of the figure. See Supplementary Table 7 for source data and P values (negative binomial).

Extended Data Table 1 | Numbers of off-target sites for gP, gM and gMH identified by Cas-OFFinder (in silico) and CIRCLE-seq (experimental)

		Sites with canonical NGG PAM							
		Number of spacer mismatches							
gRNA	Method	0	1	2	3	4	5	6	7
gP	Cas-OFFinder	1	5	41	355	1073	3347	21900	94051
	CIRCLE-seq (total WT&KI)	1	5	38	231	121	44	15	4
	CIRCLE-seq (WT mouse)	1	5	38	226	117	43	13	2
	CIRCLE-seq (KI mouse)	1	5	38	218	93	25	11	4
gM	Cas-OFFinder	1	0	0	8	77	780	8315	55093
	CIRCLE-seq (total WT&KI)	1	0	0	4	18	36	32	25
	CIRCLE-seq (WT mouse)	1	0	0	4	17	26	23	17
	CIRCLE-seq (KI mouse)	1	0	0	3	17	30	23	14
gMH	Cas-OFFinder	1	1	1	15	178	1609	10992	55363
	CIRCLE-seq (total WT&KI)	1	1	1	9	52	65	82	159
	CIRCLE-seq (WT mouse)	1	0	1	8	46	46	45	81
	CIRCLE-seq (KI mouse)	1	1	1	9	43	53	64	102
		Sites with PAM harboring single mismatch							
		Number of spacer mismatches							
gRNA	Method	0	1	2	3	4	5	6	
gP	Cas-OFFinder	0	16	370	12028	9828	20097	70495	
	CIRCLE-seq (total WT&KI)	0	14	279	2160	271	31	6	
	CIRCLE-seq (WT mouse)	0	14	272	1648	164	13	2	
	CIRCLE-seq (KI mouse)	0	14	271	1904	262	31	4	
gM	Cas-OFFinder	0	0	1	20	423	4911	34722	
	CIRCLE-seq (total WT&KI)	0	0	1	3	18	18	10	
	CIRCLE-seq (WT mouse)	0	0	0	3	15	16	8	
	CIRCLE-seq (KI mouse)	0	0	1	3	14	8	4	
gMH	Cas-OFFinder	0	0	7	80	907	8285	67109	
	CIRCLE-seq (total WT&KI)	0	0	3	19	21	9	17	
	CIRCLE-seq (WT mouse)	0	0	3	18	18	7	9	
	CIRCLE-seq (KI mouse)	0	0	3	16	9	5	13	

For sites identified by CIRCLE-seq, the total number of sites found in the wild-type and knock-in mice are listed and immediately below that are the number of sites found in each of the two mice.

Extended Data Table 2 | Off-target sites identified by CIRCLE-seq for gP, gM and gMH that exhibit single nucleotide polymorphisms or indels based on CIRCLE-seq data

Location of site	gRNA	Mouse	gRNA on-target site	Off-target sequence	Number of Mismatches	SNPs confirmed by targeted amplicon-sequencing
chr9:78832014-78832037	gM	KI	GGCTGATGAGGCCGCACATGNNG	atCaGATaAaCCaCACATGGaG	8	No
chr4:129226173-129226196	gMH	KI	CAGGTTCCATGGGATGCTCTNNG	agGGcTcacetGGATGCTCTGtG	8	N.D.
chr9:68916733-68916756	gMH	KI	CAGGTTCCATGGGATGCTCTNNG	tAGGgagagaGGGATGCTCTGaG	8	N.D.
chr3:19461541-19461564	gMH	KI	CAGGTTCCATGGGATGCTCTNNG	CAtGTaCCaAGGGATGtTCTAcG	5	N.D.
chr13:37109353-37109376	gP	KI	AGCAGCAGCGGCGGCAACAGNNG	taCAGCAGCaGCaGCAACaCGa	6	N.D.
chr6:112201818-112201841	gP	WT	AGCAGCAGCGGCGGCAACAGNNG	AGCAaCAGCaGCaGCAGCAGTaG	5	N.D.

Mismatches relative to the on-target site are shown as lower-case letters. Single nucleotide polymorphisms or indels that differ from the C57BL/6N mouse strain are shown as red-coloured letters. Targeted amplicon sequencing of sites for gM (first row) was performed with the same genomic DNA from wild-type and knock-in mice used for CIRCLE-seq experiments (data in Supplementary Table 7). N.D. = not done.

Extended Data Table 3 | Numbers of closely matched sites in the mouse genome with canonical NGG, alternate NAG and other alternate non-NGG or non-NAG PAMs for gP, gM and gMH

Guide RNA	Method for finding off-target sites	Canonical NGG PAM	NAG PAM	Non-canonical Non-NAG	Total sites
gP	Found in mouse genome by Cas-OFFinder	120773	64538	78925	264236
gP	Found by CIRCLE-seq	472	2668	263	3403
gM	Found in mouse genome by Cas-OFFinder	64274	10285	37421	111980
gM	Found by CIRCLE-seq	130	25	28	183
gMH	Found in mouse genome by Cas-OFFinder	68160	16404	79049	163613
gMH	Found by CIRCLE-seq	448	52	30	530

Sites with NAG PAM

Number of mismatches	gP sites	gMH sites	gM sites
1	0	0	0
2	14	0	0
3	331	4	0
4	11349	26	8
5	7817	220	122
6	12416	1858	1676
7	32611	14296	8479

Top, the total numbers of these sites identified by Cas-OFFinder and CIRCLE-seq.

Bottom, sites with alternate NAG PAMs identified by Cas-OFFinder are shown by the number of mismatches present in the spacer region for each gRNA.

A GROWING INSTITUTE WITH BIG AMBITIONS

Japan needs to embrace strategic research and globalization to remain at the forefront of innovation. Having joined the Okinawa Institute of Science and Technology Graduate University (OIST) at a critical phase of its growth, **PETER GRUSS** shares his vision for the institution and its contribution to Japan’s future.



After spending more than 10 years at the helm of Germany’s Max Planck Society, Peter Gruss became president and CEO of the Okinawa Institute of Science and Technology Graduate University (OIST) in January 2017. He is a molecular biologist recognized for his work in gene regulation and embryonic development, including pioneering experiments that led to the discovery of enhancers — elements in cells that amplify the activation of genes. While at the Max Planck Society, Gruss added about 200 directors of research and introduced initiatives to extend the scope of the institution’s academic pursuits. He is now tasked with nearly doubling the number of faculty and students at OIST by 2023. A founder of a biopharmaceutical company with wide experience in technology transfer and innovation, Gruss also plans to leverage his expertise to facilitate outreach to industry.

What stage of development is OIST at now?

We started out as a visionary initiative by the Japanese government to internationalize the nation’s research. Luckily, Japan hasn’t wavered in its commitment to this plan — stable funding from a single government source has contributed greatly to our growth. Since 2005, the Cabinet Office has given us roughly ¥220 billion (about US\$2 billion). Thanks to that, we’re now in a phase in which exponential growth is occurring based on the foundations we’ve laid in the past few years. In my experience, it’s unique for such an ambitious endeavour to unfold as planned. World-class institutions all look to hire the best and brightest — they all aspire to produce scientific breakthroughs that extend the limits of human knowledge. But most universities have little core funding, which makes it challenging to remain the best of the best. This is where OIST has a strong competitive advantage.

What is the secret behind OIST’s rapid growth?

From the very beginning, our research philosophy has

been to prioritize quality over quantity when hiring faculty. We also provide high funding and freedom, even at the junior level. This approach seems to be working its magic. Before the graduate university was established, we were a promotion corporation with four initial research projects, primarily in neuroscience. Shortly after inception, we began hiring in other areas of life sciences, then in physics and chemistry, and later in computer science and mathematics. Fast forward to today, and we’re producing some of the best research results in Japan. According to an analysis in Nature Index, in 2017, OIST produced the highest quality research relative to total output among Japan’s academic institutions when the playing field is leveled by institution size.

What are some of the most notable recent developments?

We’re very proud to have our first cohort of OIST students graduate earlier this year. A major upcoming goal is to increase the number of faculty from 60 to 100 by 2023 and to 200 by 2031. To accommodate this expansion, we’re currently

building our fourth laboratory building and are planning a fifth one. We also plan to solidify our strength in mathematics and computer science.

WE'RE PRODUCING SOME OF THE BEST RESEARCH RESULTS IN JAPAN

Okinawa is a unique place in Japan, both in terms of location and culture. What is the significance of having such an ambitious research institute in Okinawa?

Island sustainability and environment are important areas that we collaborate with locals on. For example, we’re involved in a project that uses new technologies to treat wastewater from Awamori distilleries and pig farms, which produce two of the most iconic staples of Okinawa. The prefectural government has been particularly generous in their support in applying this project, which also generates electricity in the process. We exert a tremendous influence on the local economy through technology transfer,

employment opportunities and our impact on education.

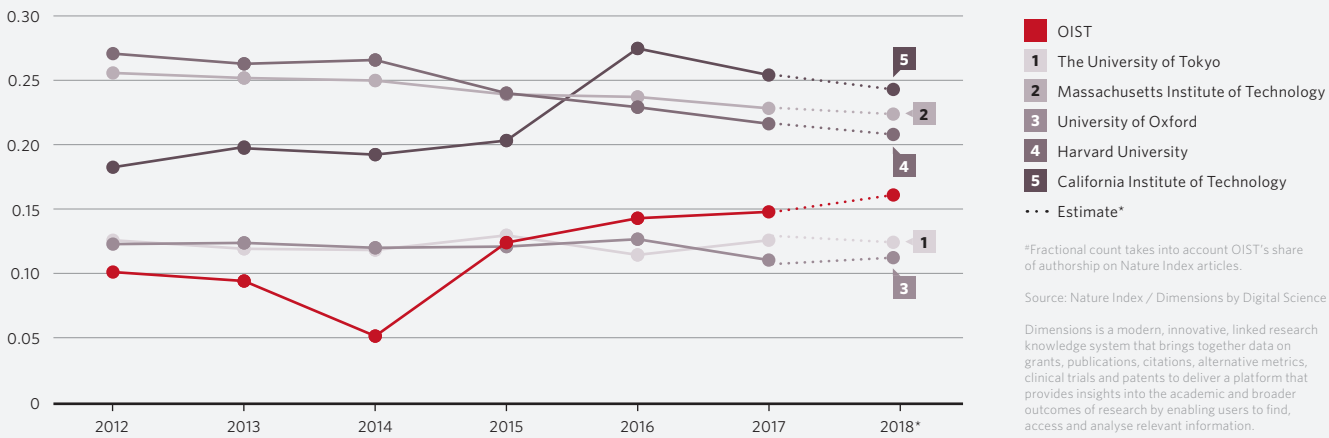
What does this mean for Japan's future?

Bringing international talent including younger scientists into Japan is vital if this country is to continue being one of the most competitive economies in the world. At OIST, we communicate in English — the universal scientific language. Being truly international helps attract academics from around the world to this campus. Ultimately, our goal is for OIST to educate and train graduates who spread this global outlook to Japanese science and beyond. I think we are making such an impact already. In fact, in some cases, OIST has now become the first port of call for Japanese postdocs and academics returning to Japan for faculty positions, because of its stable funding and facilities. In that regard, internationalizing research is also crucial for reversing the brain drain. ■



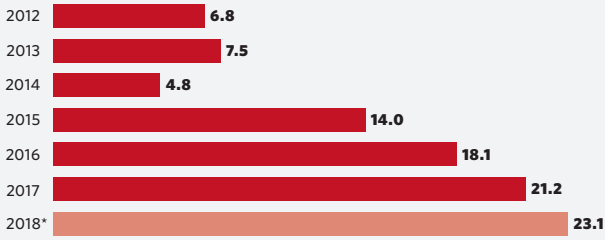
RISING BY THE NUMBERS

Normalizing the output of articles by **OIST** researchers in journals tracked by Nature Index by dividing the Fractional Count[#] (FC) by the total number of natural science articles in Dimensions allows comparison on a level playing field with much larger top institutions.



FRACTIONAL COUNT[#]

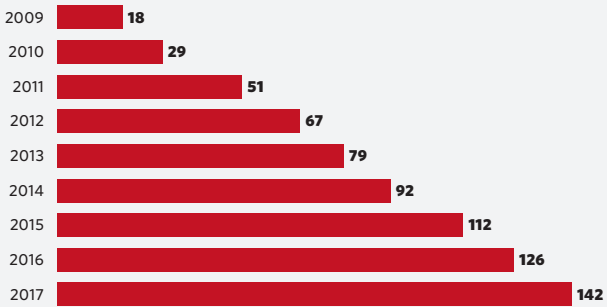
The contribution of OIST to articles in Nature Index is rising rapidly year-on-year (2012–March 2018).



*2018 is based on FC from April 2017 – March 2018 and 2017 articles from Dimensions.

OUTPUT IN DIMENSIONS

OIST’s output of articles in natural sciences is on the rise.



New star on the block rivals the heavyweights of science

OKINAWA INSTITUTE OF SCIENCE AND TECHNOLOGY GRADUATE UNIVERSITY (OIST) goes head-to-head with the big names in science.

OIST has shown rapid growth in output of articles in natural sciences since its establishment in 2011, and OIST’s output of articles published in the journals tracked by Nature Index has risen even more dramatically. As a result, the proportion of natural science articles published by OIST

researchers in the high-quality journals of Nature Index (as shown by normalized fractional count in the line graphs above) now exceeds that of Oxford University and Tokyo University, placing it at No. 1 by this measure in Japan, and OIST is fast approaching the same levels as Harvard,

MIT and Caltech. Despite OIST’s small size, its researchers covered a broad spectrum of research in life science, chemistry, marine science, physics and materials. OIST is striving to break down the walls between research fields and encourage interdisciplinary research. ■

Leveraging the huge power of microorganisms

The Institute of Microbial Chemistry demonstrates how a small research institute can **MAKE A BIG IMPACT** on both academia and the pharmaceutical industry



IMC's main campus in Shinagawa, Tokyo.

The potential of

microorganisms is almost infinite. Elucidating their metabolic mechanisms and secondary metabolites can help scientists discover novel substances that have therapeutic uses and lead to the development of vital new antibiotics.

In Japan, the Institute of Microbial Chemistry (IMC) is at the forefront of these research activities. "There are both scientific and economic difficulties to developing antibacterial drugs," says Masakatsu Shibasaki, director of the IMC. "But as long as there remains a threat to humanity, our most important mission is to produce new medicines with the aid of microorganisms."

The IMC was established in 1962 using royalties from Japan's first antibiotic — kanamycin

— discovered by renowned bacteriologist Hamao Umezawa. The parent foundation has since been managing the resulting revenue, which has enabled researchers to tackle challenging projects regardless of fluctuations in the economy. The Tokyo-based private institute, which employs about 110 researchers, has discovered about 170 bioactive compounds and brought 14 pharmaceuticals to the market, including kasugamycin, a highly effective drug against rice blast disease, and bleomycin, the world's first target-specific anticancer agent.

Shibasaki says the institute's strength lies in the balance it strikes between basic and applied research. In recent years, its reputation has grown, thanks to management reforms that have strengthened basic research and expanded the

institute's research coverage.

"If we don't produce high-quality papers, we cannot attract excellent researchers, and everything plunges into a negative spiral," explains Shibasaki, who initiated these reforms.

His reforms have fostered collaborations among microbiologists, chemists and structural biologists, and have boosted recruitment of excellent young researchers from Japan and abroad. Consequently the number of papers produced by IMC researchers has been growing steadily, leading to its ranking as the top institute in terms of the normalized weighted fractional count, an indicator of an institution's high-quality research output, in Nature Index Japan 2018. Many of these papers are from a laboratory that is synthesizing

therapeutic compounds using novel asymmetric catalysts, which encompass green chemistry principles, and another laboratory that is elucidating the molecular mechanisms of autophagy, an intracellular degradation system conserved in eukaryotes.

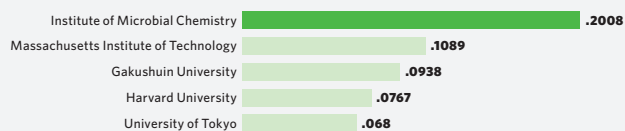
The strong fundamental research produces synergetic effects for drug discovery. The IMC focuses on developing novel treatments for diseases that many pharmaceutical companies find too costly to invest in, but for which global demand is rising rapidly — such as extensively drug-resistant tuberculosis, *Helicobacter pylori* infection, and multidrug-resistant pathogens. The IMC also maintains a rich library of natural products at a time when pharmaceutical companies are shying away from the practice.

"We cover what companies cannot afford, so our work helps realize sustainable growth in Japanese industry and science and technology," Shibasaki says. ■

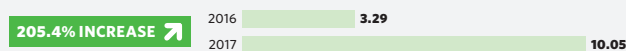
A TOP PERFORMER

IMC's output outstrips many better-known institutions

TOP INSTITUTIONS BY NORMALIZED WFC* (2012-2017)



INCREASE IN ADJUSTED FC* YEAR-ON-YEAR (2016-2017)



TOP INSTITUTION
(Nature Index 2018 Japan,
Normalized WFC* 2012-2017)



nature
INDEX 2018
JAPAN
natureindex.com

Source: Nature Index



**Microbial Chemistry
Research Foundation
Institute of Microbial
Chemistry (BIKAKEN)**
13-14-23, Kamiosaki, Shinagawa-ku
Tokyo, 141-0021, Japan
TEL: +81-3-3441-4173
FAX: +81-3-3441-7589
office@bikaken.or.jp
<http://www.bikaken.or.jp/english>

*Normalized weighted fractional count (WFC) is an indicator of an institution's high-quality research output as a proportion of total output in the natural sciences.

*Fractional count takes into account IMC's share of authorship on Nature Index articles.

Fertile ground for innovation

The Ulsan National Institute of Science and Technology (UNIST) is actively promoting technology transfer and innovation through the commercialization of its core research brands and **THE CULTIVATION OF TECH-DRIVEN VENTURE COMPANIES**

Despite a paucity of natural resources, South Korea is remarkably competitive in several international markets, including automobiles, shipbuilding, electronics and petrochemicals. Many Korean companies, including Samsung, LG, Hyundai, Kia and SK, have become global names. At the heart of the country's manufacturing boom is Ulsan, an industrial city that accounts for a sizeable portion of South Korea's gross national product and is home to the Ulsan National Institute of Science and Technology (UNIST).

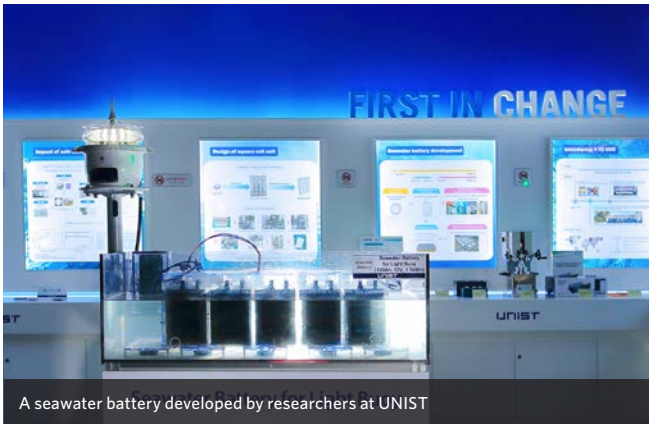
EMBRACING INNOVATION
Launched in 2009 with the vision of becoming a world-leading university to advance science and technology for the prosperity of humankind, UNIST embraces innovation. The world is currently in the throes of the fourth industrial revolution, in which three-dimensional printing, smart sensors, artificial intelligence and other emerging technologies are expected to disrupt established manufacturing practices. This transformation provides rich opportunities for innovation, and UNIST is actively seeking to cultivate new ideas.

In 2017, UNIST launched the Industry-Academia Battery

R&D Center — probably the world's largest university-operated R&D centre for rechargeable batteries. It is researching and developing small batteries used to power smartphones and mobile devices as well as large batteries for electric cars and energy storage devices. The centre is focusing on making batteries that are safer than the conventional lithium-ion batteries used in millions of laptops and cell phones today. Through the work being conducted at this centre, UNIST plans to develop rapid-charging, long-lasting rechargeable batteries.

UNIST recently announced the world's first rechargeable batteries that use the sodium ions dissolved in seawater to generate electricity and can thus be safely operated in open water with little to no maintenance. These seawater batteries are being tested for powering navigational light buoys, which guide ships away from shallow or unsafe areas. They offer several economic and environmental advantages over the lead-acid batteries that are currently used for buoys.

GARNERING GLOBAL RECOGNITION
UNIST's success stems in part from its tremendous



growth trajectory. In less than a decade, it has recruited over 300 faculty members from all over the world. The student body, which consists of about 2,500 undergraduates and 1,800 graduates, is highly international, and education at UNIST is conducted in English. Most students are scholarship recipients. Extensive support for research is another growth driver for UNIST. For example,

the campus hosts three research centres of the Institute for Basic Science (IBS). A new initiative in Korea, these research centres seek to promote excellence in basic science through the creation and support of centres that tackle important research challenges. The three centres are addressing fundamental problems in chemistry, biology and materials science, and are directed by world-renowned

UNIST AT A GLANCE

The enthusiasm and innovation of researchers at UNIST are the driving force behind the institution's remarkable achievements.

RANK AMONG SOUTH KOREAN INSTITUTIONS
(FC* 2017)

#4

RANK AMONG ASIA PACIFIC ACADEMIC INSTITUTIONS
(FC* 2017)

TOP 50

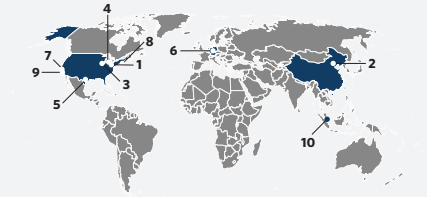
RANK AMONG GLOBAL INSTITUTIONS
(191st, FC* 2017)

TOP 200

Source: Nature Index

TOP 10 INTERNATIONAL COLLABORATORS
(FC* 2012-2017)

1.	University of Michigan, USA	15.12
2.	Chinese Academy of Sciences, China	13.18
3.	University of Illinois at Urbana-Champaign, USA	12.50
4.	Case Western Reserve University, USA	12.38
5.	The University of Texas at Austin, USA	12.33
6.	Max Planck Society, Germany	10.34
7.	Stanford University, USA	8.37
8.	Harvard University, USA	7.86
9.	University of California, Santa Barbara, USA	7.47
10.	Nanyang Technological University, Singapore	6.79



% INCREASE IN FC*
(2012-2017)

147% INCREASE



FC* AND AC (2017)



*Fractional count takes into account UNIST's share of authorship on Nature Index articles.

leaders in contemporary areas of science and technology. UNIST also provides strong support for other high-profile research centres, led by some of the world's leading scientists, including Sang Il Seok, an expert in perovskite solar cells, and Jong Bhak, who plays a leading role in international genomic research.

The university is establishing lasting collaborations with premier institutions around the globe, including the Helmholtz Association of German Research Centers, the Max Planck Society for the Advancement of Science and the Fraunhofer Society. Through these partnerships, UNIST is actively pursuing the exchange of personnel and ideas.

These activities to enhance recruitment and collaboration are reflected in UNIST's academic rankings. To list a

few, UNIST was ranked 52nd worldwide in the 2018 CWTS Leiden Ranking in terms of the proportion of the institution's publications assessed as being in the top 10% of the field (1st in South Korea for the second consecutive year) and 45th globally for citations in the Times Higher Education World University Rankings for 2018.

UNIST HAS ESTABLISHED ITSELF AS A LEADER IN RESEARCH AND EDUCATION

ENRICHING THE QUALITY OF LIFE THROUGH SCIENCE AND TECHNOLOGY

UNIST seeks to enrich of the quality of life through advances in science and technology. To this end, it has established the U-K (UNIST-Korea) research brand, which seeks

to facilitate the development and commercialization of technologies expected to enrich human life in Korea and beyond. To achieve this, UNIST will constantly explore and nurture research brands that well-represent UNIST and commercialize them to create new industries, which will in turn help to create internationally competitive innovation-led growth engines. To date, UNIST has established about 14 research brands, including Energy 4.0 Seawater Batteries, Ultra-Low Power Neuromorphic Chip (UniBrain) and Organoid 3D Bioprinting. By securing the competitiveness of these specialized technologies, UNIST plans to build K-Science, the new Korean Wave of Science. The goal is to create new innovation-led growth engines and business models for the nation and beyond.

In a relatively short time, UNIST has established itself as a leader in research and education. High-technology products that were devised and developed at UNIST are beginning to find their way into the marketplace, while UNIST's commitment to the long-term support of fundamental science is being strengthened. Students and faculty are gaining international acclaim and recognition. The steep growth trajectory of UNIST is bolstered by a dynamic framework designed for success. Like the light buoys that illuminate the sea, UNIST aims to brighten the fields of science and technology through innovation, interdisciplinarity and globalization. ■

UNIST
<https://www.unist.ac.kr>
<http://news.unist.ac.kr>
<https://www.facebook.com/unist.official>



LEADING THE WAY IN SOUTHERN CHINA

Established in 2012, **SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY** (SUSTech) has developed in leaps and bounds at its Shenzhen campus. Its international academics and lecturers conduct classes in their own style, giving every student a globally-focused and personalized education.

The Southern University of Science and Technology campus is a green oasis in the bustling metropolis of Shenzhen. Lush grass and trees surround the buildings, with brooks trickling through the grounds. Looking skyward reveals cranes throughout the university grounds as the continued expansion of facilities takes place at what locals call ‘Shenzhen speed’.

SUSTech aims to establish a world-class university, to benefit Shenzhen, Guangdong province, and China as a whole. SUSTech’s central pillars are Research, Innovation, and Entrepreneurship, in line with its historic position as a testbed for higher education

reform. SUSTech is a pioneer for new and innovative education programmes intended to cultivate and nurture talents from across the globe. The faculty and staff are drawn from far and wide, with over 90% having worked and studied overseas, and 60% having spent time at a world-renowned university.

Many of its faculty members are recipients of funding from national or provincial talent programmes, including the Thousand Talents Programme, and the Chang Jiang Scholars Programme.

As a young university, SUSTech strives for excellence and a unique educational experience for its students. By using a comprehensive

evaluation system for undergraduate admission, student evaluation is no longer solely reliant on their national college examination scores. SUSTech also considers students’ high school transcripts (accounting for 10% in admission evaluation), and applicants’ results from SUSTech’s own distinctive admission exams (accounting for 30% in evaluation).

With most of the 2018 intake graduating in the top 1% of their provinces, this system has proved highly effective in attracting students from all over China to Shenzhen.

SUSTech is not like other universities in its offering to students. Its broad focus on the lives of undergraduates

improves their preparation for life after graduation. The residential colleges provide all students with a dual-advisory system, to provide assistance in academic and personal matters. Faculty members and staff have a broad world view and training, ensuring that the students receive a global outlook when they choosing electives from a wide range of schools and departments.

SUSTech is a young and vibrant university, much like the city of Shenzhen. As the city evolves and achieves incredible milestones, SUSTech plays an integral role.

You are invited to join in and help SUSTech in its quest to become a world-class, research-oriented university. ■

Thousand Talents Global Recruitment Programme

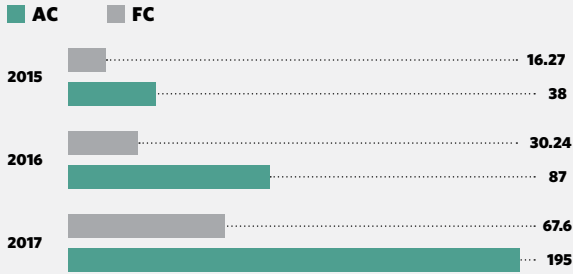
The **SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY** (SUSTech) is offering exciting opportunities for scientists and engineers in a variety of disciplines.

The Southern University of Science and Technology (SUSTech) in Shenzhen, is seeking outstanding candidates for the Thousand Talents Global Recruitment Programme sponsored by the Chinese government. Applications are invited for all major science and engineering disciplines. Successful applicants will be appointed to the faculty of SUSTech at a level commensurate with their background and experience, from tenure-track

assistant professor to tenured chair professor.

SUSTech offers a generous salary and start-up package for recipients of the Thousand Talents Global Recruitment Programme. Benefits include: a competitive starting salary; a living subsidy of 2.75 million RMB (Thousand Young Talents) or 4.5 million RMB (Thousand Talents and Foreign Thousand Talents) over three to five years; a start-up fund of up to 12 million RMB; principal investigator and tenure-track

NATURE INDEX SCORES FOR 2015 TO 2017



A **315%** increase in FC from 2015 to 2017.

#32 AMONG CHINESE UNIVERSITIES
Based on Nature Index fractional count data from January 2017 to December 2017



talents@sustc.edu.cn
www.sustc.edu.cn/en
+86-755-88010968
88010945





上海科技大学
ShanghaiTech University

Opportunities to shine at ShanghaiTech University

ShanghaiTech University is a young and dynamic higher education institution aiming for high-quality research and global influence. To address challenges faced by China and the world, it seeks innovative solutions in energy, materials, environment, human health, data science, artificial intelligence (AI), and electrical engineering. An integral part of the Zhangjiang Comprehensive National Science Center, the university is now leading several frontier research projects at large-scale facilities. For more information, please visit: www.shanghaitech.edu.cn.

We are now seeking talented researchers for multiple faculty positions at all ranks in the following fields:

School of Physical Science and Technology: energy, systems materials, photon and condensed light states, material biology, environmental science and engineering

School of Life Science and Technology: molecular and cell biology, structural biology, neuroscience, immunology, stem cells and regenerative medicine, systems biology and biological data, molecular imaging, biomedical engineering

School of Information Science and Technology: artificial intelligence, electrical engineering, computer science and technology, information engineering, communication engineering, VR, statistics

School of Entrepreneurship and Management: economics, finance, management, marketing, strategy and entrepreneurship

Shanghai Institute for Advanced Immunochemical Studies: immune antibodies

iHuman Institute: bio-imaging, biology, chemistry, computational biology, AI/ML

Institute of Mathematical Sciences: pure mathematics, theory of computing, applied mathematics

Successful applicants will have a doctoral degree, and are expected to establish a record for independent, internationally recognized research, supervise students and teach high-quality courses.

ShanghaiTech University will offer attractive compensation packages, including:

initial research support package: reasonable start-up funds, research associates and post-doctoral fellows, laboratory space to meet research needs

compensation and benefits: highly competitive salary commensurate with experience and academic accomplishments, a comprehensive benefit package

subsidized housing: on-campus, 80/100/120 m² faculty apartments available at low rent for tenure and tenure-track faculty

relocation & travel allowance: reimbursement of expenses for household relocation and family's one-way travel

family assistance: support with children's education; affiliated kindergarten, primary and middle schools are under construction

To apply: using this format, please submit a cover letter (Firstname_Lastname_Cover_Letter.pdf), a research plan (Firstname_Lastname_Research_Plan.pdf), and a CV (Firstname_Lastname_CV.pdf) to shanghaitechuniversity@gmail.com.



CAREERS

ONLINE Career resources from our scientific community nature.com/careers

SOCIAL MEDIA Follow us on Twitter at twitter.com/naturejobs

GOT A STORY? Contact us at naturecareerseditor@nature.com

ALICE MOLLON/GETTY



GENDER GAP

She persisted

Six women talk to Nature about navigating gender bias in academia.

There is plenty of evidence to show that gender bias plays out against women in academic hiring, tenure and promotion, as well as in teaching evaluations. The unconscious, ingrained nature of gender bias and discrimination can make these barriers pervasive and hard to prove.

Combined with widespread sexual harassment in both laboratories and fieldwork, these phenomena can drive young women away from academic careers.

Yet many universities, and some nations, are making progress in advancing women's careers. For individuals, however, it can be difficult. *Nature* spoke to six senior academic female scientists about their advice and strategies for navigating gender bias.

POLLY ARNOLD Seek senior allies

Crum Brown chair of chemistry, University of Edinburgh, UK, and founder of SciSisters, a network for senior women in STEM in Scotland.

When I was a junior academic, the university recognized that female faculty members were winning international awards, but they were not being promoted internally at the same rate as their male colleagues.

So they sent the junior female faculty

members on a leadership course — even though, according to external metrics, we were doing brilliantly. We all talked, and realized that we didn't need training to convert us into men. We needed more support as women who are leaders — such as opportunities to meet other female scientists and discuss our career plans. The more diversity you have on your team, the better your results.

Until that course, we hadn't met many women in the same position. It was lovely. The chance of being a lone woman in a senior position in science, technology, engineering and mathematics (STEM) fields is high, and we didn't have a way to find each other to share difficult stories.

SciSisters launched in 2017 as a Google map through which women can find each other ►

► locally. People can also look at the map to find an expert, or younger women can look for mentors.

To embed changes that counteract gender bias, you need to have buy-in not only from senior women, but also from senior male colleagues who will listen. If those people are in the room, then I feel supported, and not like the 'screaming feminist'. My department thinks about inclusion all the time. We are unafraid of calling each other out, probably because we all have coffee together frequently.

I use killer data to counter bias. I'll say, "Have you read the research that shows that women are perceived, by both men and women, to have spoken more in a conversation, even when they have spoken less?"

HOLLY DUNSWORTH

Model expectations

Associate professor of biological anthropology, University of Rhode Island, Kingston.

There is a bias among some university students, who are less likely to call female professors 'Doctor'. I changed my e-mail alias to 'Dunsworth' and my e-mail signature to 'Dr. H. Dunsworth', because I would get e-mails from students that started: "Hey Holly!". I want the same status and respect that my male colleagues receive.

I wear an academic robe to lectures. Students call me wacky, but it neutralizes judgement of my wardrobe and is one less thing for students to evaluate negatively about me. I thought it would take students' attention off my body and my clothing choices, and it did.

But I'm not trying to deny who I am. I have brought my son to class to model that this is also how professors can look. It's relevant to my research that adult female primates will, more often than not, have a kid hanging off them as they go about their work.

I explain to students why I do this. We talk about earning respect on the basis of how you look, and about data that show students hold higher expectations for female professors than they do for male professors.

Perhaps because women are seen as more nurturing, students expect to be cared for, and I get inappropriately personal e-mails from them that feel overwhelming. It's not a problem unique to women, but after conversations with male colleagues, I'd say it's worse for women.

So, to set a boundary, I have guidelines in my course syllabus that say, "You will receive some notifications about this course via e-mail. However, e-mailing Dr. Dunsworth is not part of this course." The guidelines encourage them instead to come to my office and talk face to face: "Let's be Stone Age humans together." There's been a drastic reduction in the number of pointless and inappropriate e-mails that I receive.

HANAN MALKAWI

Keep pushing

Vice-president for science engagement, Royal Scientific Society; microbiologist at Yarmouk University, Amman, Jordan.

When I finished my bachelor's degree in biology at Yarmouk University in 1981, I had the highest scores in the department. I was offered a scholarship to get my master's degree and PhD at a university outside of Jordan. I chose Washington State University (WSU) in Pullman, but encountered resistance from my extended family. I had grown up in a strict Muslim family, who felt it would be dangerous for a young, single woman to live alone in a Western country.

Even my undergraduate department chair suggested that maybe I should wait a few years, and get married first, before studying abroad. Despite the fact that my uncles were against my going abroad, my father insisted and said, "She has to go." I did go. I eventually met a man from Jordan who was also studying for his PhD at WSU, and we got married. After I earned my PhD, I returned to teach, and continued my career at Yarmouk.

I became dean for research and graduate studies, then vice-president for research and international relations. At one point, I was the only woman sitting on the dean's council, and I was responsible for a lot of committees.

It was a challenge. I perceived that because I was young and female, not all of them accepted me immediately. But my science background helped me to be wise. Sometimes, I treated all my dean colleagues and faculty members as if I were their sister, their peer, and not their boss. I would say, "These are the responsibilities, and we'll distribute them among all of us, including

me." At other times, I would have to be tough.

I worked alongside them for hours, helping and mentoring. After a few months, I gained their respect and their confidence.

Today, young, single, female graduate students still face the dilemma that I did about studying abroad. A few years ago, I knew a student in Jordan who was offered a scholarship to do her PhD in the United States. I went three times to convince her family. Her father finally said he would go with her for one semester. When she finished her PhD, she came back to Jordan as an assistant professor to be a role model for other young women here.

JESSICA MEIR

Build confidence

Comparative physiologist and astronaut, NASA Johnson Space Center, Houston, Texas.

At NASA, a lot of women don't dress in a very feminine fashion. In technical and operational fields, people tend to view women who look feminine as less competent. But if everyone dresses the same, we're never going to change anything. People should wear what they want to wear. For me, I can still be an astronaut and wear a skirt. If people see that, then they will not equate femininity with being operationally incompetent.

I've seen cultural differences in gender bias at other space programmes. Although I don't think the intentions were malicious, I've heard comments or jokes that would be sexist in my culture. Often, I just ignore it, but sometimes I try to turn it around with a quick-witted or sarcastic reply to make them see how absurd the comment is or to make them think a little bit.

My first battle with unconscious gender



Astronaut Jessica Meir at skills training for NASA.

JAMES BLAIR/NASA



Rebecca Calisi Rodríguez works with a rock dove (*Columba livia*) in her lab's aviary.

GREGORY URQUIAGA/UC DAVIS

bias was experiencing impostor syndrome in graduate school. This was something I saw in almost all of my female colleagues, but only very rarely among the men. But during my PhD defence at the Scripps Institution of Oceanography in La Jolla, California, it was like a light bulb went on: "This is a ridiculous amount of work I've done. I do deserve this. I know more about oxygen depletion in diving penguins and seals than anyone." That gave me the confidence I really needed.

As women, we often find it hard to act or portray ourselves confidently. Even if we are more confident, we might use language or expressions that come across as not being confident.

Take that step back to gain perspective on your work. It's easy to get lost in the daily details. But look at it from the bigger-picture level, and appreciate it as if somebody else were doing it.

Make sure you are doing what you are passionate about. Your mannerisms will convey enthusiasm for it. And people will believe that you are capable and knowledgeable. Do something that gives you a sense of purpose. These things will make you a better advocate for yourself, and people will listen to and notice you because confidence comes out naturally.

As women, we need to be willing to promote ourselves, because it has an implication for our own success. We are not going to get that seat at the table if we don't grab it.

REBECCA CALISI RODRÍGUEZ Speak up

Assistant professor of reproductive biology, University of California, Davis.

When I started my position in 2015, there was a university-wide meeting for new faculty

members — on a Saturday. My husband, who was a postdoc at that time, also attended. We brought the kids with us and gave the 5-year-old the iPad, and I wore the baby carrier with my sleeping 3-month-old son.

We went around the table and everyone gave their name and department. My husband gave his name and, as I was about to speak, the man to my right started instead. They skipped me even though I was sitting right at the table.

I navigated this by being vocal. I spoke up: "Hey, I was skipped over. I'm faculty in the neurobiology department." The room just paused.

I decided right then that I was not going to be overlooked. But being overlooked as a woman at meetings happens all the time. Or I'll say something that will be interrupted by a man, then restated and credited to him.

I want the credit for my ideas, but I also don't want to rub my colleagues up the wrong way. And as women, if we're really forceful, then we're seen as too aggressive. It's a fine line. My female colleagues and I are being strategic to help amplify each other's voices. When another woman speaks up in a meeting, we restate it and credit that woman specifically.

I was given terrible advice early on to view other women as adversaries and to outshine them. But the truth is, we rise by lifting each other. And it's not specific to women. Men can also amplify the voices of female colleagues to create equity.

At conferences, everyone needs to increase their tolerance for tiny interruptions by tiny people or breastfeeding mothers. By supporting women's needs and a culture of inclusion, we'll retain more women.

I tell young women that these problems are everywhere, not just in academia. That might sound negative, but I don't want them to leave academia thinking things will be better elsewhere.

Still, things are changing. There's a huge community of women, men and non-binary

people — so many allies out there who are becoming more vocal. Look for these groups. Look for supportive mentors. I'm going to be here to help make that change.

JAELYN EBERLE Do your homework

Vertebrate palaeontologist; director of the museum and field-studies graduate programme at the University of Colorado Boulder.

You need an excellent mentor of any gender. Get to know people, see how they react to situations, and find somebody whom you have considerable respect for and get along with. They should be senior enough to have experience and clout at your institution.

Stay clear of unhappy people who have a jaded opinion of the university. You need someone who is willing to go to bat for you and has a positive, forward-thinking attitude. If you don't get assigned a mentor, or one doesn't work out, go to your department chair with three suggestions of people you'd like to have as your mentor.

Another piece of advice for new faculty members: negotiate at the very beginning. That's hard, but really important. You've got to do your homework to learn what someone with your education and credentials normally receives as salary and lab space.

Negotiate at those big times of getting tenure and promotion, too. Find out what other professors in your department were promised when they made tenure. Ask to be paid equitably with others at your ranking.

Running expeditions in remote places such as the Arctic for weeks or months at a time taught me quickly which personalities work and which don't in field teams. It's important to have gender diversity on every team.

You've also got to be capable of all the things the team will need to do. I purchased a shotgun and learnt how to use it, because all Arctic expeditions require a gun in camp for protection against polar bears.

Whether it is setting up the communications radio, knowing outdoor first aid or reading the weather to guide a helicopter coming in — if you are knowledgeable and feel confident about your skills out there, you'll make good decisions. Equip all your people, regardless of gender, with the proper skills, equipment and knowledge that they need, too.

In the end, we strive for people to be treated as people — not according to their gender identity.

When I'm discussing the evolution of mammals after dinosaur extinction, does it matter that I'm female? It shouldn't. ■

INTERVIEWS BY KENDALL POWELL

These interviews have been edited for clarity and length.

THE 133RD LIVE PODCAST OF THE GOURMANDO RESISTANCE

A taste of freedom.

BY BETH CATO

Claudia had been a devotee of the Gourmando Resistance Podcast since episode 20. She knew what to do as hostess. She knew she may not have much time to do it. Her battery-operated stove burner was assembled, the large pot atop it half full of oil at approximately 180°C. Her dough was mixed. Half of it was loaded into an old plastic device that, oddly enough, was called a gun, although this device extruded dough in various shapes and sizes, depending on the disc that was loaded into the end.

And, most importantly for the podcast, she had on her full synth suit, complete with a gustatory sheath on her tongue. But for the first time, she was going to output data, not receive.

Through the overlay on her enhanced contacts, she stared at the camera mounted above her makeshift kitchen. With the movement of her eyes, she signalled the programme to go live.

"Greetings, fellow Gourmet Commandos! Welcome to the 133rd Live Podcast. I'm Claudia." She showed off the plastic gun with its dough-filled tube. "I'm making a dessert my grandma remembered from her childhood: churros."

She bit back the urge to babble nervously. She had to get cooking. Nutrition enforcers spied on the Gourmando forums online, and they might be watching her even now. Her face was mostly bare — part of the defiance of the broadcast, and the risk. If she was identified via facial or retinal recognition software, officers could be at her door in minutes.

It was impossible to know how many 'casters were busted, but very few hosted multiple episodes.

Claudia pulled the trigger on the gun to release a fat, star-shaped tube of dough into the oil. It baffled her that people cooked food in the old days using a method that smelt so awful. Things were so different before the famines and strict caloric monitoring, before daily rations of government-issued AllFood loaves in five flavours.

"This churro dough is simple, just flour, sugar, salt, butter, water and eggs." She squirted more dough into the oil. "Thank you to everyone who sent ingredients and tools."



Gourmandos had contributed supplies, as they always did. Hundreds, maybe thousands of people, tuned in for these broadcasts to experience illicit foods of yore. Everyone stayed anonymous until they volunteered to take a turn at the camera. Each time Claudia had checked her assigned drop-box locations around the city in recent weeks, she'd wondered if she'd find ingredients, kitchen implements or a trap.

"The churros are turning brown fast," Claudia said. "Oh! I almost forgot." She set down the gun and touched the dough still in the bowl, letting the sensors in her fingertips share the data with everyone else who wore synth suits. Those people already knew her armpits were getting downright swampy.

In a sudden fit of bravado, she pinched off some dough and brought it near her mouth. She could imagine the horror of many of her viewers. Some might even be ripping the sensor sheaths from their tongues, repulsed by the idea of experiencing raw ingredients, especially eggs.

But she remembered the giddy delight she'd felt when a previous podcaster had shared raw chocolate-chip cookie dough. Claudia had thought the dough had tasted even better than the fresh-baked cookies.

The churro dough wasn't anywhere near as delightful. She almost spat it out, but after a few quick chews, she swallowed.

"I think the churros are done now." She used a slotted spoon to transfer them to a towel on a plate. Some churros were curved, whereas others were fairly straight. Strangely enough, the oil smelt good now. She breathed in deeply, allowing the foreign scent to drift through the filters over

her nose and to her fellow foodies.

Claudia glanced at the clock. Not even ten minutes had passed. She hesitated, tempted to start more dough in the oil, but no. She needed to finish the first batch. She couldn't miss the chance to experience what her grandmother had told her about in a creaky voice so full of yearning.

A quick tap confirmed a churro was just cool enough to touch. She tossed it into a prepared bowl mounded with cinnamon and sugar, and rolled the churro around for an immersive baptism. The incredible sweet and spicy scent made her eyes and mouth water.

Claudia brought the completed churro to her lips. After a lifetime of living on AllFood loaves, of vicariously tasting the forbidden through the podcast, she was going to ingest contraband calories for herself.

She bit into the tip of the churro. The coating of sugar and cinnamon dazzled her tongue as her teeth crunched through the outer ridges to find an interior that was soft, chewy and delightfully hot. She took another bite and wept. Salty tears joined the divine flavours in her mouth.

"I'm eating a churro, Abuelita," she whispered.

A heavy knock shuddered through the door.

No. No. No —

An instant later, the digital locks were overridden. Officers in armoured suits burst in, guns in hand — and these guns weren't loaded with dough. She froze, overwhelmed. She'd read so many theories about what happened from here: hidden tribunals, imprisonment, prosecution. That rebellious casters were probably stripped of their enhancements, forevermore isolated from the expansive digital world.

But she had experienced churros, and so much more. This had been worth it.

She shoved the rest of the churro in her mouth. "Keep cooking, Gourmandos!" she shouted. Crumbs sprayed from her mouth as she held high a fist that sparkled with sugar. Her head was slammed into the table a second later. Through dizzying pain, she swallowed, and smiled. ■

Beth Cato resides in Arizona. She's the author of both *The Clockwork Dagger* duology and *Blood of Earth* trilogy published by Harper Voyager. Her website is BethCato.com.

ILLUSTRATION BY JACEY

THE WORLD AT THEIR FEET

These newcomers are making their mark in science across the disciplines.

From cutting the cost of solar electricity to reducing the risk of ovarian cancer, the 11 early- to mid-career scientists profiled here are emerging as leaders in their fields.

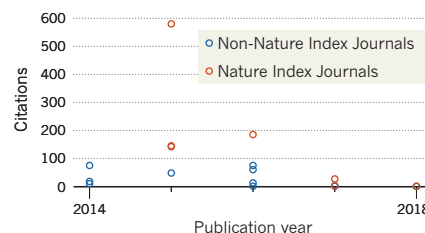
They stood out from among 500 scientists assessed using the power of the Nature Index and the League of Scholars Whole-of-Web (WoW) rankings. They are bringing fresh ideas in a range of disciplines, from cognitive neuroscience to geology, and condensed matter physics. Their initiative, curiosity and flexibility have given them an edge in a competitive research environment.

The analysis included active researchers who have published at least one paper in the 82 index journals in 2017, and whose first scientific paper appeared less than 20 years ago, with some even emerging on the scholarly scene in the past six years.

The profiled scientists have shown year-on-year citation growth, and scored exceptionally in the WoW ranking, which identifies the most influential researchers using an algorithm similar to Google's PageRank. It considers factors such as the quality of a scientist's output, links to industry, and co-authorship networks.

RESEARCH RECORD

Dane deQuilettes's publication history. Each dot represents one paper, with citation figures correct as of 3 August 2018. Red dots represent papers published in one of the 82 journals tracked by the Nature Index. Some dots overlap.



DANE deQUILLETES, 28

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

DEFECT DETECTIVE

A chemist seeks the right material to bring cheap, solar electricity to those without access.

Materials scientist **Dane deQuilettes** hopes to help transform the world's energy systems, especially for people who don't have reliable access to electricity. The postdoc at MIT is using his expertise in the properties of promising materials called perovskites to make it happen.

The performance of perovskites in solar cells rivals conventional silicon, and they promise to be much cheaper to make. While costs of silicon solar cells, which supply 1.7%

of the world's electricity, have come down, they are not likely to fall low enough to make an impact for the 1.2 billion people globally who don't have access to grid electricity.

GridEdge Solar is an MIT project, led by deQuilettes, to evaluate different light-weight, flexible photovoltaic materials, and, in a few years, build a pilot line to manufacture them. He thinks perovskites, which can be made into ink-like solutions and printed on rolls of material, similar to newsprint, are a strong contender. The project is funded by the Indian philanthropic organization Tata Trusts.

deQuilettes, who has a background in chemistry, started working on perovskites during his PhD at the University of Washington. Fascinated by their particular crystalline structure, he has studied how single misplaced atoms in perovskites can degrade their properties and inhibit their performance in a solar cell. Before his work, it wasn't clear why there was so much variation in the quality of perovskite solar cells. deQuilettes revealed the answer in the varying atomic alignment in different regions of the crystalline materials. "Once we understood where the defects were, we could work on a design strategy to remove them and make the performance of the device more uniform," he says.

"I'm interested in the fundamental physics of how light interacts with materials, but I always come back to, 'What's the purpose?'" says deQuilettes. **KATHERINE BOURZAC**

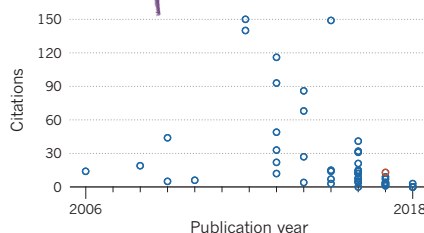
HEARTFELT REASONER

A cognitive neuroscientist reveals how the body moves the mind.

Sarah Garfinkel studies how beats of the heart, and our awareness of its rhythms, can influence everything from anxiety levels and emotional learning, to sleep quality and racial bias. She's now one of the world's foremost experts on the health consequences of 'interoception', the felt sense of one's internal organ activities.

As a postdoc at the University of Michigan, Garfinkel studied memory recall among veterans of the Iraq and Afghanistan wars who were suffering from post-traumatic stress disorder. Her study focussed mostly on the brain, but a curious observation led Garfinkel to wonder about the role the heart might play in emotional processing. Why was it that the heartbeats of some veterans remained steady while in the brain scanner reliving their traumatic experiences, whereas the hearts of others raced frantically?

Working with Hugo Critchley, a neuropsychiatrist at the Brighton and Sussex Medical School, she revealed a disconnect between how good people think they are at detecting their own heartbeats and their true



SARAH GARFINKEL, 38

BRIGHTON AND SUSSEX MEDICAL SCHOOL

accuracy. Garfinkel then showed why that incongruence matters, reporting in 2016 that the less people with autism knew their own hearts, the greater their anxiety.

In another study, she and Critchley found that poor interoceptive awareness among people with depression or anxiety was associated with deficient sleep quality. "This now gives us a target for intervention," Garfinkel says. "We want to train people to have better precision over their bodily signals."

Garfinkel is now co-leading one of the first clinical trials of an interoception-directed therapy, evaluating whether a computer training module can help people with autism become more in tune with their heartbeats and thereby reduce anxiety.

Critchley says psychologists had been aware of interoception for more than a century, but its clinical importance had largely been overlooked until Garfinkel's work. "She reinvigorated this whole field," he says. **ELIE DOLGIN**

ELECTRON MICROSCOPIST

A condensed-matter physicist peers deep into materials for industrial applications.

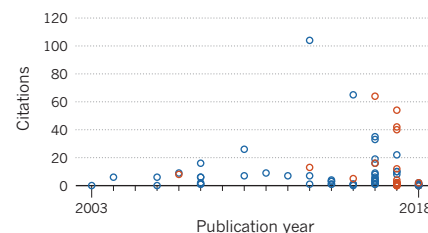
Binghui Ge made his mark with an answer to Richard Feynman's 1959 challenge: "Is there no way to make the electron microscope more powerful?" Invented in 1931, transmission electron microscopy (TEM) vastly

improved the resolution of conventional optical microscopes by using a beam of electrons, instead of light, to reveal nanometre-sized structural features. But Feynman urged researchers to improve the resolution by a hundred times. Drawing on imaging theory, Ge and a team at the Chinese Academy of Sciences's Institute of Physics developed a method for obtaining structural information less than a nanometre in size, using conventional TEM. They provided the first analytical expression of image distortions that arise in samples of greater thickness and used that to observe individual atoms.

Ge has moved on to the application of TEM to reveal the microstructures of catalysts and thermoelectric materials in unprecedented detail. Recently, he has been exploring the microstructure of thermoelectric materials at multiple scales, with a view to improving their efficiency in heating, cooling and generating power. These materials could be used to convert the wasted two-thirds of heat produced by vehicle engines to electricity.

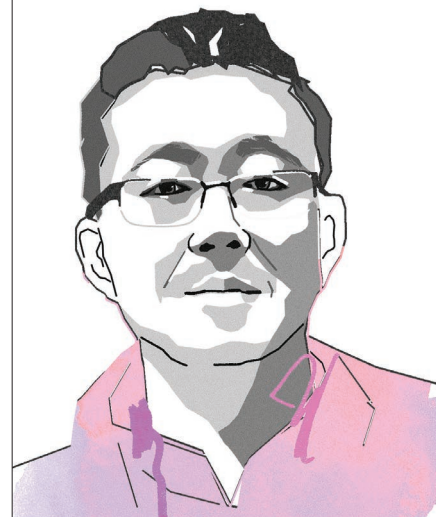
Ge wants to explore cryo-electron microscopy, a TEM method allowing scientists to image three-dimensional biological molecules without destroying samples. Ge wants to know whether this can improve imaging of non-biological materials that are also susceptible to beam damage, such as metal-organic frameworks, widely used in catalysts.

CATHERINE ARMITAGE



BINGHUI GE, 40

INSTITUTE OF PHYSICS, CHINESE ACADEMY OF SCIENCES



FROST SURVEYOR

A physical geographer digs deep in frozen soils to fill gaps in maps.

To **Gustaf Hugelius**, white space on a map presents a challenge. The modern-day explorer, who as a boy hiked every summer in northern Scandinavia, looks for peatland and permafrost areas of the Arctic and sub-Arctic regions where soils have not been analysed. The gaps in the maps have taken him as far as Siberia, Greenland, and northern Canada.

Arctic peatland and permafrost represent 25% of the Earth's carbon sink. Global warming is expected to thaw these frozen grounds, releasing carbon dioxide and methane into the atmosphere and accelerating climate change. Yet their potentially crucial role in the carbon–climate feedback cycle is poorly understood, partly due to large gaps in the data.

Hugelius has pioneered the use of high-resolution satellite imagery, calibrated with field samples, to show that northern soils, due to seasonal freezing and thawing, are far more variable and complex in composition than the comparatively simple 'back-yard' soil types used in climate modelling. Reflecting the diversity in soil composition and, correspondingly, in the decomposition process that releases greenhouse gases, could improve the accuracy of climate models.

The field research is "very demanding", he says. There are permits to secure; there are the logistics of getting people, equipment and supplies to some of the world's most remote regions for long stays; and the sheer difficulty of drilling into ice-hard ground for samples. Not to mention the polar bear risk. "You need to be aware of the bears, both for their safety and for ours," he says.

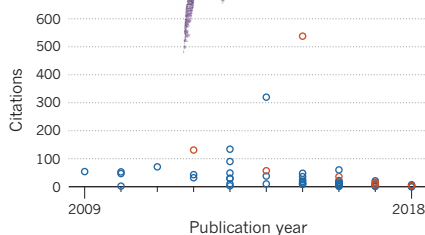
Poor data sharing has long been a problem for soil research, he says. As manager of the Northern Circumpolar Soil Carbon Database, a dataset of organic carbon stored in soils of the region he studies, Hugelius is helping to fix that, also working with climate modellers so they can better account for uncertainty in the models. **CA**

READY-TO-WEAR DESIGNER

A system engineer develops stick-on sensors to monitor health.

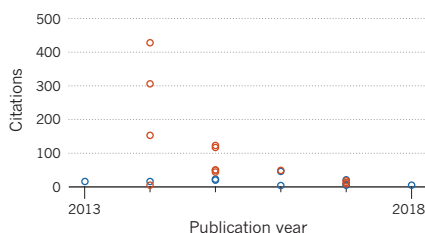
One day, stretchy sensors applied to the skin, like temporary tattoos, will monitor vital signs, predicts **Jaemin Kim**. If things go according to his plan, sophisticated and squishy electronics will give robots a sense of touch.

Kim, a system engineer from South Korea, became interested in wearable electronics



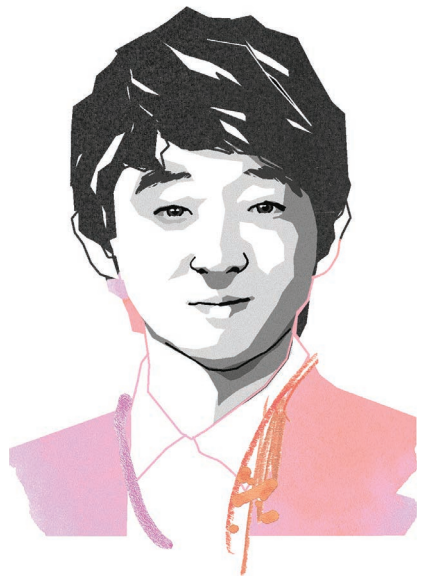
GUSTAF HUGELIUS, 38

STOCKHOLM UNIVERSITY



JAEMIN KIM, 31

STANFORD UNIVERSITY



during his master's degree at the Korea Advanced Institute of Science and Technology in Daejeon. "I tried to measure emotion by measuring goosebumps," he says. The best way to sense the changes, he learned, was not with typical, rigid electronics, but pliable ones that move with the skin.

Throughout his PhD at Seoul National University, Kim developed sensors and systems for stretchy devices, including a sticker-like heart monitor that uses a printed heart-shaped display to tell the wearer whether her electrocardiogram is normal (red heart) or unhealthy (blue).

He has worked on a soft, curved image sensor that could one day become the basis for a retinal implant for blind people. But, he says, a lot of difficult engineering work is needed to make these promising devices practical.

The flexible electronics in his publications have to be attached to bulky external electronics and power sources to work.

In November 2017, he moved to California to work as a postdoctoral researcher at Stanford University with Zhenan Bao, one of the leading chemists designing intrinsically stretchy electronic materials. At the Bao lab, Kim is focussing on using these materials to make standalone medical and touch sensors that don't have to be tethered to a bulky microcontroller to analyse the data they gather.

Kim says he's torn about whether to continue in academia or take an industry role. He wants to see this research translated to products. "I want to make something that works for everyone," he says. **KB**

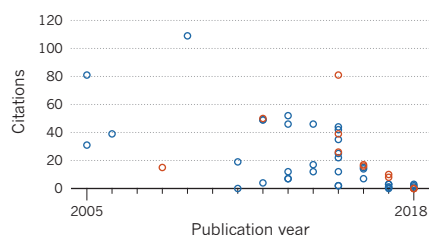
MINI-MOLECULE MANIPULATOR

An organic chemist creates hydrogels made of cost-effective, self-assembling proteins.

Antimicrobial peptides, made of chains of amino acids, are the body's first line of defence against invading pathogens. These proteins have the geometric quality of chirality, which means they cannot be overlaid on their mirror images, just as a left-handed glove doesn't fit on a right hand.

Their chirality also determines their biological activity — a quality that pharmaceutical manufacturers have long utilized to create drugs with sought-after properties. The heartburn pill, Nexium, for example, is made from a left-handed molecule. Its predecessor drug, Prilosec, also made by Astra-Zeneca, included the left and right hands of the molecule pair.

These molecules generally consist of long strings of hundreds of amino acids. Organic chemist, **Silvia Marchesan**, of the University of Trieste in Italy, has a more refined and cost-effective approach. She works with



SILVIA MARCHESAN, 39

UNIVERSITY OF TRIESTE



short peptides, only three amino acids long, and switches the chirality of the individual amino acids. “It’s like putting a right-hand finger on a left hand to see what kind of hand we get, and how this new hand behaves differently,” she says.

Marchesan has used this technique to make tripeptides that self-assemble into water-based gels that have intrinsic antimicrobial properties and are biocompatible. The supramolecular structure of the hydrogels allows the potential to switch functions on and off, making them useful as enzyme substitutes, scaffolding in the repair of body tissue, and for the sustained delivery of drugs.

In 2013, Marchesan co-authored a paper, which described combining the self-assembling tripeptides with a common antibiotic. The resulting hydrogel continuously released the drug over six days.

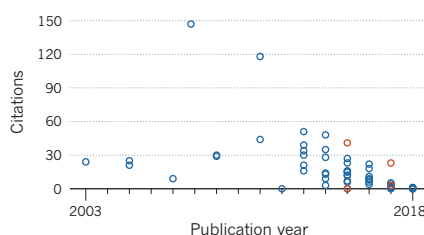
A paper that made the August 2018 cover of *Chem* describes why Marchesan’s tweaked tripeptides behave differently from their real-world analogues and how the process holds consistent across the scale from single molecule to macroscopic hydrogels, which is important for large-scale production. The next step, says Marchesan, is to refine the lab process into one that can be scaled up cheaply and sustainably. **CA**

CANCER SLEUTH

A marine biologist turns molecular epidemiologist to tackle ovarian cancer risk.

It’s hard to study marine biology from land-locked Vermont. So, as an undergraduate, **Melissa Merritt** left to spend a semester investigating the reproductive abilities of corals along the Great Barrier Reef in northeast Australia. Within a few years, though, Merritt had to give up ocean studies due to problems with sea sickness, and went into cancer research. Her grandmother had ovarian cancer and Merritt decided to devote herself to elucidating risk factors and genetic drivers of the gynaecological disease.

For her PhD and postdocs, Merritt trained in cancer epidemiology and molecular biology at institutions across Australia, the



MELISSA MERRITT, 41

UNIVERSITY OF HAWAII CANCER CENTER



United States and the United Kingdom — gaining experience and a dual scientific background that made her “very adept at formulating important research questions”, says molecular epidemiologist Marc Gunter, a former mentor from Imperial College London.

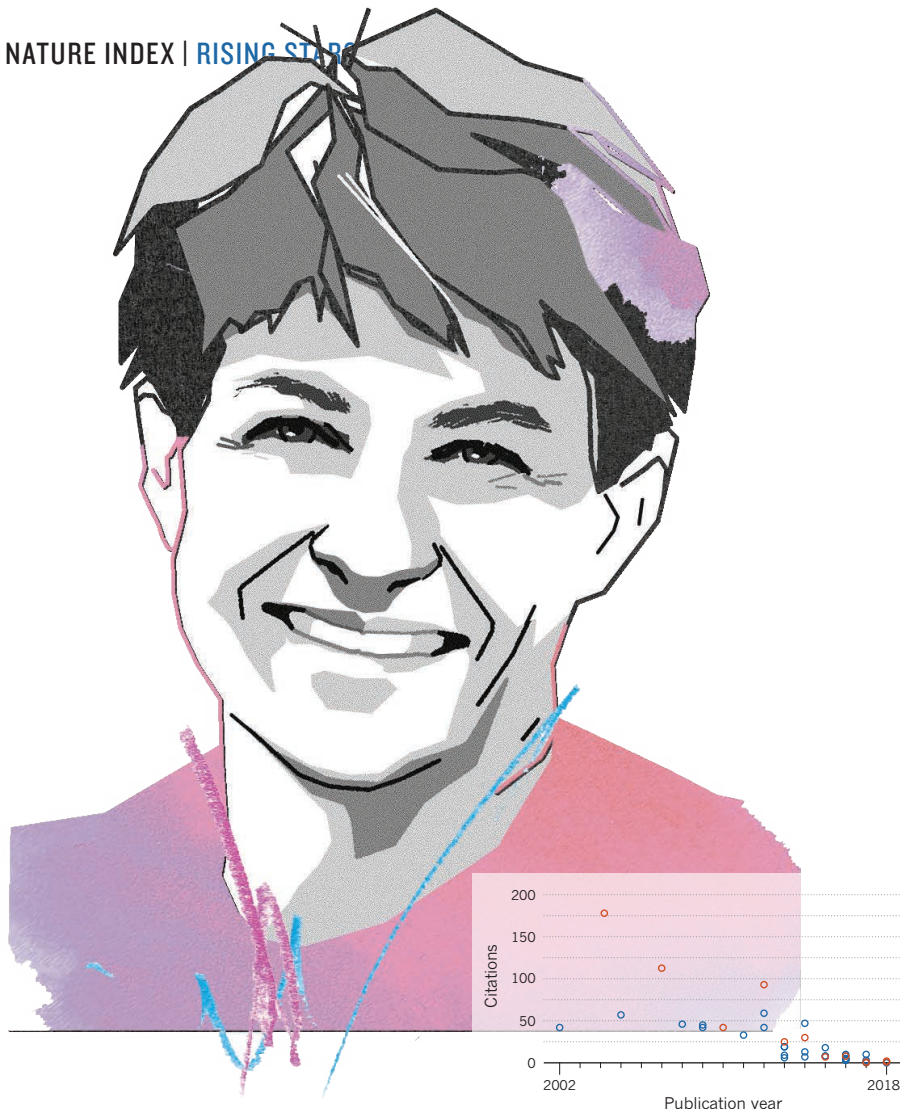
In 2015, for example, Merritt helped develop a methodological approach for evaluating the risk of dietary factors in cancer. It is now being used to study the links between foods and tumours of all kinds, and has also shown that coffee intake helps lower a woman’s risk of developing endometrial cancer.

In what Merritt considers her most “significant finding”, she showed in a recent study that women with locally invasive ovarian cancer have approximately a 30% lower risk of dying from the disease if, after their diagnosis, they take aspirin or a nonsteroidal anti-inflammatory drug like ibuprofen.

Now a faculty member of the University of Hawaii Cancer Center in Honolulu, Merritt has secured funding to explore whether hormone-altering chemicals found in many products affect a woman’s chances of developing endometrial cancer — a project that will harness Merritt’s skills in basic lab science and epidemiological data analysis. **ED**

SOURCE FOR ALL GRAPHS: NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

ILLUSTRATIONS BY PADDY MILLS



CLIMATE ROCKER

A geologist studies how changes in climate wear away mountains and riverbeds.

By geological standards, the timescales that **Taylor Schildgen** studies are short. She is trying to determine how climate transforms the Earth's surface over thousands of years. Her work uses a new technique to date landforms, by measuring the presence of rare isotopes known as cosmogenic nuclides. The proportion of these isotopes in rock samples allows geologists to estimate their age and rate of change over millennia.

In 2017, Schildgen published a paper, with postdoc, Stefanie Tofelde, considering the influence of global climate variations on river terraces in the Argentinian Andes. The researchers examined 100,000-year cycles of glaciation and interglaciation. They found that during colder and wetter periods, the increased flow of water and sediments cut deep slits in the valley. The basins filled back up with sediment during drier, warmer periods. Nearby river channels closer to the mountains responded similarly, but over cycles of 21,000 years.

These studies could offer clues about what parts of the landscape will be

TAYLOR SCHILDGEN, 40

**GFZ GERMAN RESEARCH CENTRE FOR
GEOSCIENCES/UNIVERSITY OF POTSDAM**

sensitive to the sudden changes in climate we are experiencing, says Schildgen. The results have “big implications for things like flood hazards and water management,” she says.

Occasionally, Schildgen's work takes her further back in time. In July 2018, she co-authored a paper in *Nature*, refuting assumptions about the link between global climate and erosion. She established that there was insufficient evidence to suggest that the onset of glacial–interglacial cycles several million years ago accelerated erosion in alpine regions.

Schildgen traces her fascination for rocks back to a vacation to Yellowstone National Park as a teenager. After completing her PhD in geology at MIT in 2008, she took a postdoctoral position in Germany due to the shortage of academic positions in the United States. “Part of me feels guilt for having left the US because I received such good training there,” she says. “But the reality of my situation is that I have been extremely well supported in Germany.” **SMRITI MALLAPATY**

MATERIALS MAESTRO

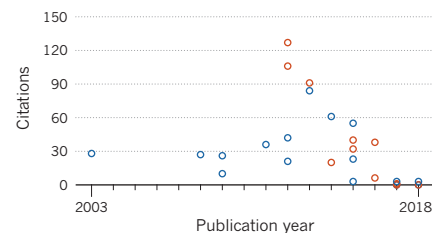
A computational physicist delves into the complex structures of organic materials.

Due to their abundance and low cost, organic molecules make excellent building blocks for flexible nanoelectronic devices. But achieving the properties needed to replace inorganic semiconductors, such as silicon, demands an intimate understanding of how the molecules interact at the nanoscale.

Sahar Sharifzadeh employs computational modelling to find intuitive ways of describing these molecular interactions, which could facilitate the development of organic materials whose electrical conductivity can be controlled with exquisite precision — a characteristic of semiconductors.

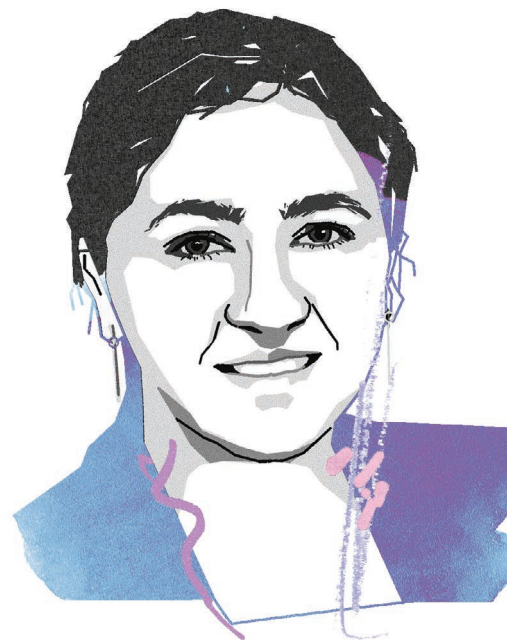
Sharifzadeh has used this theoretical approach to show that the arrangement of molecules in an organic crystalline material strongly influences how it responds electronically to light — an insight that could improve the efficiency and longevity of organic solar cells. Her computational analyses have also been applied to understanding the structures of promising inorganic materials.

In 2017, she simulated free-standing,



SAHAR SHARIFZADEH, 36

BOSTON UNIVERSITY



one-atom-thick sheets of boron, known as borophene. Researchers have only recently grown the two-dimensional material on silver, but have not been able to isolate it in its freestanding form. Like its close relative graphene, borophene promises exceptional properties, including strength, flexibility and electrical conductivity. Sharifzadeh's calculations predict that borophene's optical and electronic properties could be precisely adjusted by stretching or compressing the material.

Sharifzadeh, whose family moved to the United States from Iran when she was eight, started out studying electrical engineering and computer science at the University of California, Berkeley, but the physics courses she took in that degree captured her imagination. Her PhD on condensed matter physics at Princeton University led to a two-year post-doctoral project on nanoscale materials at the US Department of Energy's Lawrence Berkeley National Laboratory in California. She set up her lab at Boston University in 2014.

She is modelling not just the behaviour of molecules, but that of the high-achieving scientist who is also the mother of a one-year-old. Aware that many young women fear the two roles are incompatible, she's happy to be seen doing both. **CA**

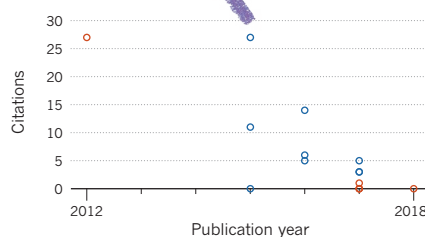
MOTION TRACKER

A biophysicist develops microscopes to peer into living tissue samples.

As a summer intern at the private foundation Brain and Spine Institute (ICM) in Paris, **Olivier Thouvenin** helped develop an imaging technique for monitoring the neural circuitry in zebrafish larvae. But the method lacked the spatial resolution to tease apart connections between neurons at the single-cell level, which frustrated the young biophysics master's student.

So for his PhD at the Langevin Institute — two kilometres away on the edge of Paris's famed Botanical Gardens — Thouvenin worked on improving an existing high-resolution tissue-imaging tool called full-field optical coherence tomography (OCT). He added a dynamic time element to the otherwise static picture-taker, and with that, Thouvenin says, "we could see things that were moving inside the sample". Thouvenin used his upgraded OCT technique to track the movement and metabolism of subcellular organelles and other structures inside living retina tissue from mice and monkeys.

Last year, he returned to the same ICM lab as a postdoc, with the microscope he'd designed in tow. Working with neuroscientist, Claire Wyart, he showed in



OLIVIER THOUVENIN, 27

LANGEVIN INSTITUTE

as-yet-unpublished research that cerebrospinal fluid in zebrafish larvae moves in both directions through the backbone, as in humans. Flow disturbance could result in spinal curvature defects, in fish and people.

About to start his own lab back at the Langevin Institute, Thouvenin is also thinking about how to make a bigger societal impact through research. "He's a very generous and open-minded person," Wyart remarks. Currently, most commercial OCT instruments used for diagnosing eye diseases and other health problems cost €30,000 (US\$34,800) or more, which makes them inaccessible for many hospitals in developing countries. Thouvenin hopes to bring the price down by using three-dimensional printing, cheap optical parts and a smartphone as a camera. "I'm building a prototype," he adds, one that should cost less than €1,000. **ED**

FOREST MODELLER

An ecologist applies mathematical modelling to forest management.

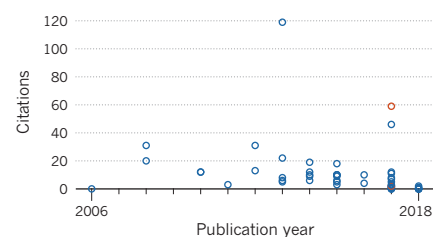
Giorgio Vacchiano started his career establishing the impact of climate change on forests, but is now finding ways of using forests to mitigate climate change.

His PhD thesis linked the high mortality of a hardy tree species in Italy's southwestern Alps to drought. The finding set the foundation for an influential paper, published

in 2013, presenting strong evidence of the effects of climate change on forest coverage over a span of 20 years. A co-authored paper in *Nature Climate Change* in 2017 reviewed more than 600 studies and found overwhelming evidence of the role of climate change in the increasing frequency and magnitude of fires, droughts and pests.

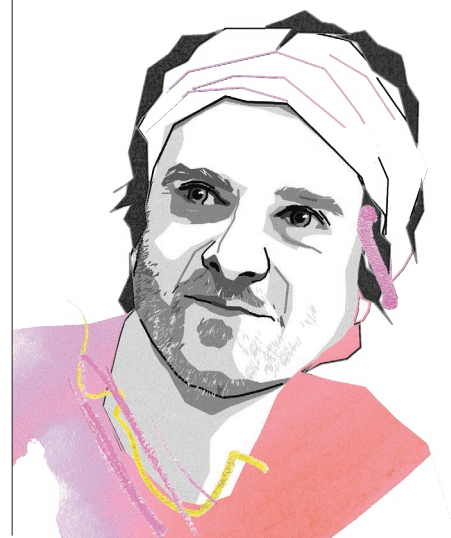
Vacchiano's interest in conservation prompted him to study forestry, but he now believes in managing forests for multiple purposes. In Italy, he has pioneered the use of modelling for forest management. Using tools to simulate dynamics under different conditions allows estimation of, for example, how quickly trees will grow after a thinning, or how many trees are needed to stop rocks falling from a slope. These tools offer forest managers a more accurate and reproducible basis for decisions, he says. Vacchiano spent 15 months working on forest modelling with the European Commission before landing his current position at the University of Milan in 2017. He considers himself lucky. "There are a lot of brilliant young researchers who can't find research work in this country."

Vacchiano's current research focusses on optimizing forest management to mitigate climate change, including harvesting timber to replace more carbon-intensive materials such as concrete for building and fossil fuels for energy. **CA**



GIORGIO VACCHIANO, 38

UNIVERSITY OF MILAN



CHALLENGER STATES

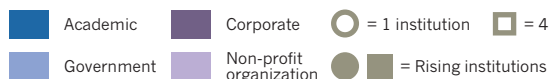
These six countries have experienced the highest absolute and percentage increases in their contribution to the Nature Index since 2015. While China is making waves among the traditional scientific powers, the other five nations are disrupting lower-tiered research strongholds.

QUALITY GROWTH

Assessed on their contribution to high-quality research in the natural sciences, all six countries have upped their pace of production since 2015. Iran stands out for its 30.7% increase in fractional count (FC), from 66.87 in 2015 to 87.43 in 2017. China is on a scale of its own, accelerating 22.6% from an FC of 7,412.96 in 2015 to 9,088.90 in 2017, just under half that of the world leader, the United States.

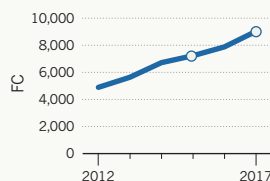
SECTOR STARS

Not all sectors appear equal, and not all sectors rise together. In Austria, in the past three years, a higher proportion of corporate institutions (23/28) than academic institutions (18/26) have increased their article counts in the Nature Index. Norway's growth has been more evenly spread across all four sectors.



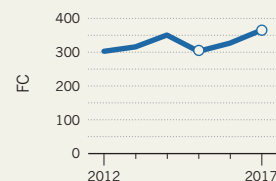
CHINA

R&D SPENDING
(% GDP, 2015): 2.06%
RESEARCHERS
(FTE, 2015): 1,619,027
TOP RISING INSTITUTIONS (2017):
1. University of Chinese Academy of Sciences (FC: 255.65)
2. Tsinghua University (FC: 353.40)
3. Shanghai Jiao Tong University (FC: 166.39)
Read more on page S27

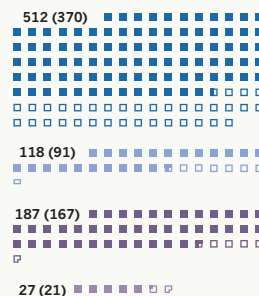


AUSTRIA

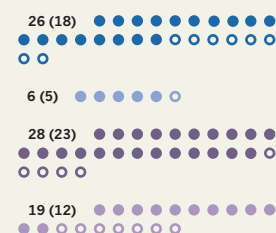
R&D SPENDING
(% GDP, 2015): 3.05%
RESEARCHERS
(FTE, 2015): 43,562
TOP RISING INSTITUTIONS (2017):
1. University of Vienna (FC: 65.42)
2. Institute of Science and Technology Austria (FC: 24.37)
3. Austrian Academy of Sciences (FC: 36.03)
Read more on page S31



Total institutions: 844 (649 rising)

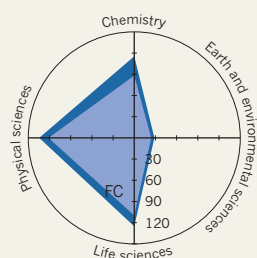
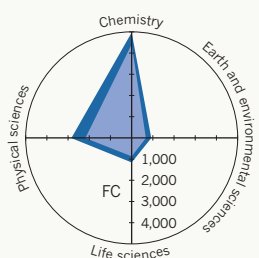


Total institutions: 79 (58 rising)



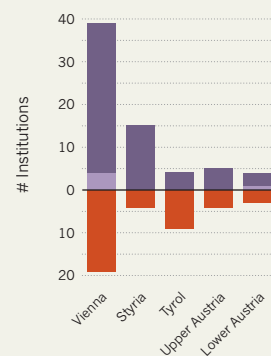
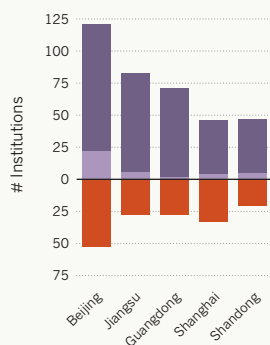
SUBJECT STRENGTHS

National acceleration is often driven by research specializations. Iran and Brazil excel in the physical sciences, and Norway in Earth and environmental sciences. This is not the case in the Czech Republic. While scientists in the country favour chemistry, their fractional count (FC) has swelled in the three other fields.



REGIONAL CLIMBERS

This graph shows the five administrative provinces, counties or states with the most institutions in each country. In China, Beijing remains the centre of knowledge production in the natural sciences, but Jiangsu, Guangdong and Shandong have a higher proportion of rising institutions. Other regions that rise high above the tide are Styria in Austria, Oslo in Norway and Tehran in Iran. Rising institutions had a higher article count in 2017 than in 2015.



*Merged with Nord-Trøndelag in January 2018 to form Trøndelag



NORWAY

R&D SPENDING

(% GDP, 2015): 1.93%

RESEARCHERS

(FTE, 2015): 30,632

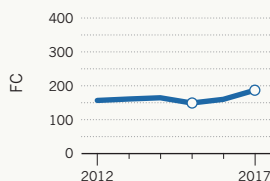
TOP RISING INSTITUTIONS (2017):

1. University of Oslo (FC: 65.58)

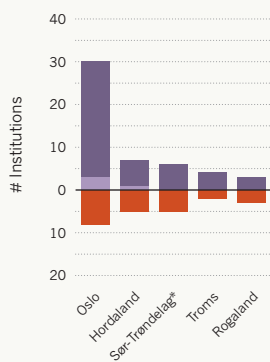
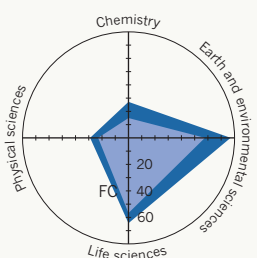
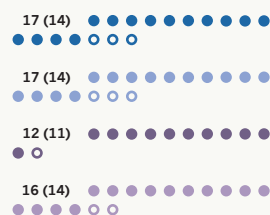
Read more on page S29

2. Norwegian University of Life Sciences (FC: 8.62)

3. University of Tromsø - The Arctic University of Norway (FC: 15.97)



Total institutions: 62 (53 rising)



CZECH REPUBLIC

R&D SPENDING

(% GDP, 2015): 1.93%

RESEARCHERS

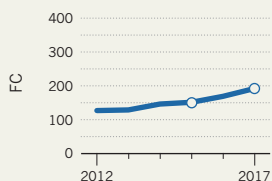
(FTE, 2015): 38,081

TOP RISING INSTITUTIONS (2017):

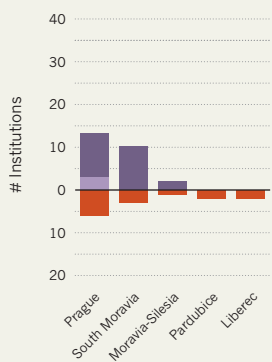
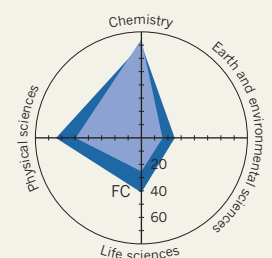
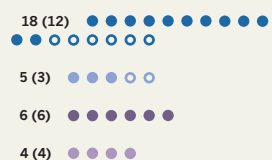
1. Czech Academy of Sciences (FC: 86.84)

2. Masaryk University (FC: 21.97)

3. Silesian University in Opava (FC: 5.24)



Total institutions: 33 (25 rising)



BRAZIL

R&D SPENDING

(% GDP, 2015): 1.28%

RESEARCHERS

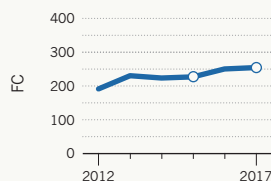
(FTE, 2014): 183,853

TOP RISING INSTITUTIONS (2017):

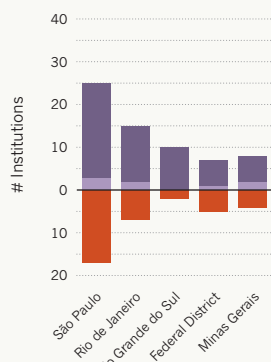
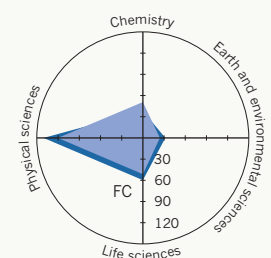
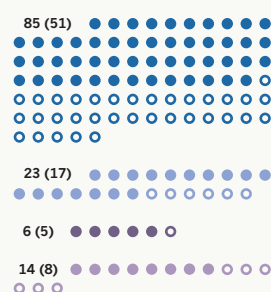
1. University of São Paulo (FC: 56.35)

2. Ministry of Science, Technology, Innovation and Communication (FC: 17.16)

3. Federal University of Minas Gerais (FC: 12.90)



Total institutions: 128 (81 rising)



IRAN

R&D SPENDING

(% GDP, 2013): 0.25%

RESEARCHERS

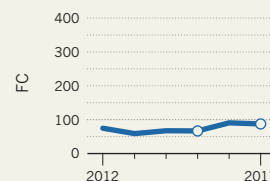
(FTE, 2013): 51,961

TOP RISING INSTITUTIONS (2017):

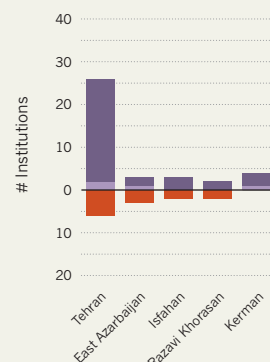
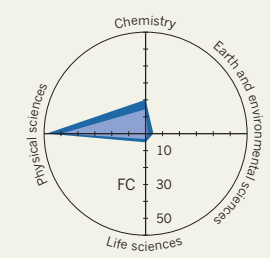
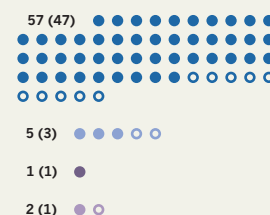
1. University of Tabriz (FC: 3.60)

2. Yazd University (FC: 3.09)

3. Damghan University (FC: 3.27)



Total institutions: 65 (52 rising)



SOURCE: UNESCO/OECD

SOURCE FOR ALL GRAPHS: NATURE INDEX



COMMENT
DAVID LANGLEY
THERINA THERON

DISCOVERY RELIES ON STRONG SUPPORT STAFF

A lack of trained administrators is holding African scientists back.

The global research funding system is becoming increasingly complex and competitive. Scientists need to demonstrate quality, relevance, impact and innovation, while meeting the highest standards of integrity and ethics, managing intellectual property issues and publicizing their work.

To help them succeed in this demanding environment, scientists in the global north have something that those in the global south lack: comparatively well-resourced support structures with trained research management and administration staff, who assist with planning, developing, managing and sustaining their research pursuits. Higher-education institutions in the global south would benefit from similar investments in strong, multi-skilled research support.

Many regions are struggling to compete with the global scientific powers. Africa, for example, has some 700 universities serving more than 1.2 billion people. The region has the lowest investment in research and development, the lowest number of researchers per capita and a comparatively low, albeit growing, share of global scientific publications.

Just as the most talented athletes need good coaches to take them to stardom, gifted scientists in the region need well-trained support staff to help them shine.

IN SHORT SUPPLY

In 2014, the 55 countries that make up the African Union adopted a 10-year strategy for science, technology and innovation. They recognized the critical role of science in the socio-economic development and growth of the continent, and stressed the importance of turning universities into centres of excellence.

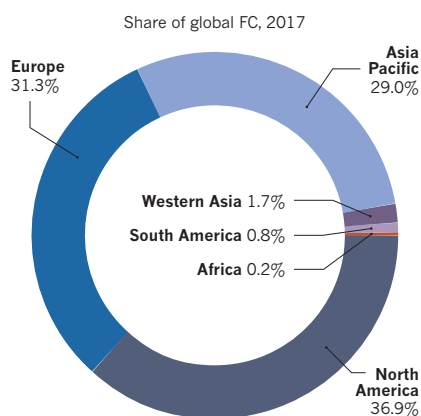
But progress has been slow. The top country in terms of spending on R&D, South Africa, is still below 0.8% of GDP and showed slightly decreasing spending over the past five years.

Universities in Africa are increasingly dependent on international sources of funding for research. Funders in the global north are also recognizing the value of collaborating with researchers in low- and middle-income countries. They offer geographical advantages

in certain fields, talented researchers, access to important data sets, knowledge of the developing world context, and insight into new markets. Among others, the United Kingdom Newton Fund, the United States National Institutes of Health, and the European Union Horizon 2020, have committed millions of dollars in grant funding for collaborative research with Africa and other regions in the global south.

RESEARCH REPRESENTATION

In 2017, institutions in Africa contributed to 0.2% of the paper authorship in the 82 journals that make up the Nature Index. Their contribution to high-quality research in the natural sciences, measured by fractional count (FC), was far exceeded by institutions in North America, Europe and Asia Pacific.



To capitalize on these opportunities, universities need to invest in skilled research support teams. Well-trained professionals can instil confidence in potential partners that research funds will be spent efficiently and responsibly.

But, such professionals are in short supply on the continent. The majority of universities in Africa typically rely on senior professors to provide part-time guidance on research administration, while expecting them to maintain teaching and research obligations. There is therefore a general lack of capacity in African universities to support and develop research, and no obvious career pathway or professional

qualification to allow for the recruitment and development of research management and administration personnel.

ON COURSE

Over the past decade, several initiatives have sought to create a research support system in Africa and other parts of the developing world. These have been supported by the Association of Commonwealth Universities, as well as international funding bodies, including the World Health Organisation and the Wellcome Trust. Existing professional associations for research and innovation managers, such as the Southern African Research and Innovation Management Association (SARIMA) and WARIMA in West Africa, have been strengthened through capacity development interventions, with new associations created in Central (CARIMA) and East Africa (EARIMA) and the Caribbean (CabRIMA).

In 2017, a consortium of institutions led by Stellenbosch University in South Africa, and co-funded by the EU Erasmus+ programme, launched the first professional academic qualification programme for research management and administration in the global south. The project, known in short as StoRM, will develop a post-graduate diploma course and a master's degree curriculum, as well as a mechanism for formal recognition of professionals in the field. It will also promote exchange between administrators in Europe and southern Africa.

Universities also need to invest more in this area to support their scientific stars. For example, they could allocate indirect costs recovered by grant income to support structures. Many universities in the global south are developing strategies to become research intensive. Providing optimal research support will help create thriving and sustainable scientific communities, and improve innovation and impact from research in the developing world. ■

David Langley is chief partnerships officer at *New Model in Technology and Engineering, UK*. **Therina Theron** is senior director of research and innovation at *Stellenbosch University, South Africa*.

SOURCE: NATURE INDEX

JOHANN THERON PHOTOGRAPHY; GABRIELLA KARNEY PHOTOGRAPHY



Excavations led by Griffith University at Leang Bulu Bettue, a cave on Sulawesi island, Indonesia.

MOVERS AND SHAKERS

These 16 institutions were selected from among the most improved institutions in the Nature Index between 2015 and 2017. Some showed exceptional absolute and relative growth in their overall contribution to the papers in the journals tracked by the index, measured by fractional count (FC), while others excelled in a specific subject category. Chinese institutions make up more than half of the top 100 rising stars, far exceeding the 20 from the United States, and four each from Germany and the Netherlands (see online tables).

GRIFFITH UNIVERSITY

AUSTRALIA | 2015 FC: 12.8 | 2017 FC: 24.09

Spread across five campuses in Queensland, Griffith University boasts a strong commitment to sustainability, as befits an institution near the Great Barrier Reef.

The university achieved an 88.1% increase in its fractional count between 2015 and 2017, making it the fastest-rising Australian institution in the Nature Index. Ned Pankhurst, a marine biologist and senior deputy vice chancellor at Griffith, attributes its recent success to two strategic investment programmes: one focussed on specific research areas, including environmental science, water science and climate change response, and a 2020 plan to become one of the most influential universities in the Asia-Pacific region.

Those programmes together have seen an

additional Au\$60 million (US\$43.1 million) in funding since 2010. Overall, Griffith has 1,446 full-time researchers and had a total research budget in 2017 of Au\$280 million.

Hui Jun Zhao, an analytic chemist, and director of the Centre for Clean Environment and Energy is among the university's top contributors to the index. He recently co-authored a paper in water research of a self-cleaning system for monitoring raw sewage in real time.

Griffith is also a founding member of the International WaterCentre, a partnership between four Australian universities that engages communities, practitioners and academics in finding solutions to complex water management problems.

In 2016, Griffith University launched the Australian Research Centre for Human Evolution, which contributed to the recent discovery of the oldest fossil of modern humans outside Africa. **BIANCA NOGRADY**

SOUTHEAST UNIVERSITY

CHINA | 2015 FC: 37.97 | 2017 FC: 75.60

Southeast University (SEU) has long been a science hub in China. It was once part of Nanjing University, but split away in 1952.

Today, its historical strengths in chemistry and physics have powered its growth in the Nature Index, with its output nearly doubling over the past three years. Publications have included research on atom-thin memory storage devices and graphene-based electronic tattoos that can monitor heart rates.

One of the university's most widely read papers in 2017, however, was in the life sciences: a fruit fly study that investigated the neuronal circuitry that drives flies to either sleep or have sex. Neurobiologist, Yufeng Pan, and his colleagues at the MOE Key Laboratory for Developmental Genes and

Human Disease observed that sleep deprivation caused male fruit flies to more often sleep rather than mate, whereas female fruit flies seemed impervious to fatigue.

Pan credits recruitment of a large number of foreign researchers — including through China's Thousand Talents Plan — for SEU's growth. Though research in life sciences only began at SEU in the past decade, he thinks the recruitment trend will continue with the recent creation of a joint institute on neuron morphology with the United States-based Allen Institute. The university has 925 doctoral supervisors. **MARK ZASTROW**

UNIVERSITY OF CHINESE ACADEMY OF SCIENCES

CHINA | 2015 FC: 101.95 | 2017 FC: 255.65

The Chinese Academy of Sciences spans a network of more than 100 research institutes across China — and its affiliated university matches its scale. Headquartered in Beijing, the University of Chinese Academy of Sciences (UCAS) sprawls across four campuses with five satellite branches in other cities, including Shanghai and Chengdu. It counts more than 15,000 instructors and 7,210 doctoral supervisors on staff, with more than 45,000 graduate students.

UCAS is the latest iteration of China's first graduate university in science, founded in 1978. Still regarded as China's best graduate school, UCAS began to admit undergraduates in 2014.

The university has shown remarkable growth in its output of high-quality research: from 2015 to 2017, its fractional count increased 151% from 101.95 to 255.65, the largest rise of any institution globally. The pace of growth is sustained across all fields tracked by the index, and contrasts with a 3.4% decline in the output of CAS over the same period. **MZ**

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

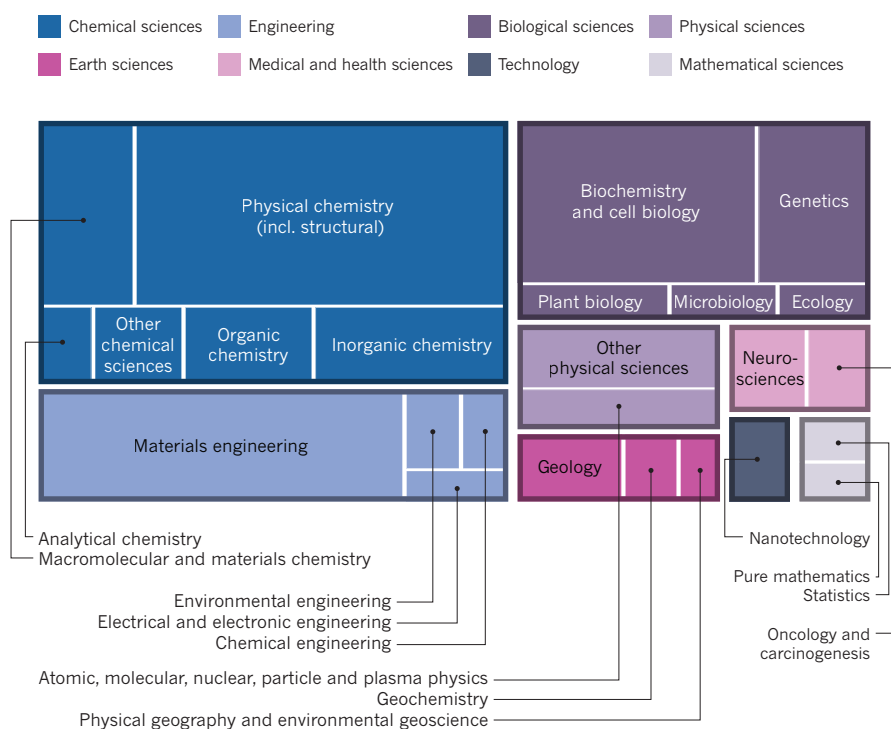
UNITED STATES | 2015 FC: 31.55 | 2017 FC: 47.02

The National Center for Atmospheric Research (NCAR) in Boulder, Colorado, was founded in 1960 to assist researchers in their studies of atmospheric and climate science. The aim was to provide faculty members with resources no single university could afford. Those resources now include the NCAR-Wyoming Supercomputing Center, the Mauna Loa Solar Observatory in Hawaii, and two aircraft for geoscience studies.

These facilities have contributed to NCAR's rise in prominence in the Nature

SPECIALTY FIELDS

Researchers at the University of Chinese Academy of Sciences (UCAS) — the top rising star in the Nature Index — are prolific producers of natural science papers in physical chemistry, materials engineering and biochemistry and cell biology. This analysis is based on the field categories assigned to UCAS articles published between 2015 and 2017 in the 82 journals tracked by the Index. Boxes are sized by number of articles.



Index, with a 58.8% increase in its contribution to journals in Earth and environmental sciences between 2015 and 2017.

Growing interest in collaborative research has brought NCAR to the fore, says Gerald Meehl, who heads its climate change research section and is among its top authors in the index. He also attributes the institute's success to a high percentage of eager young project scientists who are pretty active," he says.

More than half of NCAR's funding — US\$173 million in 2017 — comes from the United States National Science Foundation, with additional funding from other federal agencies interested in weather and geosciences, including the National Aeronautics and Space Administration, the National Oceanic and Atmospheric Administration, and the departments of defence and energy. NCAR has approximately 494 scientists, postdocs, and engineers on staff. **NEIL SAVAGE**

RUSSIAN ACADEMY OF SCIENCES

RUSSIA | 2015 FC: 155.38 | 2017 FC: 170.84

Founded in 1724 by decree of Peter the Great, the Russian Academy of Sciences (RAS) is the world's oldest scientific research network, and represented the height of Russian science through the era of the Soviet

Union. But after the Cold War, the institution struggled with budget cuts as state funds shifted towards universities. In 2013, parliament implemented reforms to the institution, seeking to bring it under control of a federal agency that reports directly to President Vladimir Putin. Critics said the move threatened the academy's independence.

However, RAS has rebounded in the Nature Index — especially in chemistry, where its output has risen by more than 50%, from a fractional count of 44.21 in 2015 to 66.73 in 2017. In October 2017, an all-RAS team demonstrated a more sustainable 3D printer that uses a material derived from cellulose, the bulk material in plant walls.

Valery Rubakov, a theoretical physicist at the RAS Institute for Nuclear Research in Moscow, says the value of RAS is more recognized today than five years ago. But he warns that if the architects of the 2013 reform continue to exert control, "the system of RAS institutes is in danger of complete destruction." RAS's budget in 2013 was 60 billion roubles (US\$887 million), and it employed 46,955 researchers in 2016. **MZ**

WESTERN UNIVERSITY

CANADA | 2015 FC: 32.95 | 2017 FC: 46.12

The 140-year-old Western University in

Ontario, Canada, has 12 faculties with 1,396 full-time staff and a research budget of more than Ca\$225 million (US\$172 million) in 2017. The university's fractional count has increased by 40% in the Nature Index from 2015 to 2017. Its output in the physical sciences has doubled, with noticeable growth in chemistry and the life sciences. Researchers during this period have contributed to numerous papers on emerging materials, astrophysics, and the brain regions controlling behaviours such as paying attention.

John Capone, vice-president (research), attributes the rise to the university's significant investments in areas such as neuroscience, materials science, wind engineering and medical research. In 2013, for example, the university invested more than Ca\$20 million in clusters of research excellence focussed on cognitive neuroscience and musculoskeletal health.

Western University has also secured private and public partnerships, particularly in areas of research with higher equipment and infrastructure costs. Overall, it has invested more than Ca\$400 million in infrastructure, including biomedical imaging and high-performance computing facilities at The Brain and Mind Institute. **BN**

UNIVERSITY OF CALIFORNIA, IRVINE

UNITED STATES | 2015 FC: 125.68 | 2017 FC: 162.5

University of California, Irvine (UCI) is North America's top rising star in the Nature

Index, achieving 29.3% growth in its fractional count over the past three years, with increases in all fields.

Founded in 1965, UCI is one of the younger schools in the state's university system, but has come into its own thanks to an emphasis on building research facilities with advanced equipment, such as the recently opened Irvine Materials Research Institute that includes a state-of-the-art transmission electron microscope.

UCI has also invested in recruitment. "We attract really top-notch faculty," says Soroosh Sorooshian, director of the Center for Hydro-meteorology and Remote Sensing in UCI's engineering school. The university's strategic plan, released in 2016, set a target to recruit 250 new faculty members by 2021, welcoming 57 in 2017 alone. That's a continuation of a hiring plan to recruit a few senior professors and several junior faculty, particularly if they span different departments. Recent high-profile hires include tissue bioengineer Kyriacos Athanasiou, and cognitive scientist, Zygmunt Pizlo, who uses computational modelling to simulate how humans perceive three-dimensional shapes from 2D images.

The university has more than 2,100 research faculty. Its funding for fiscal year 2017 was US\$378 million. **NS**

TU DORTMUND UNIVERSITY

GERMANY | 2015 WFC: 23.40 | 2017 WFC: 41.82

Located in Germany's industrial heartland, TU Dortmund University has made fast friends in the 50 years since its establishment. Collaborations have ensured its position among Germany's top rising stars in the Nature Index, almost doubling its fractional count (FC) since 2015, with even faster growth in chemistry from an FC of 11.86 to 25.59 in 2017. The partnerships have been facilitated by a combination of strong researcher networks, university strategy and a changing grant environment.

In 2012, TU Dortmund joined a €28 million (US\$32 million) initiative, funded by the German Research Foundation, to explore the science of solvents. Chemists, physicists and engineers from 50 research groups in seven institutions are part of the fundamental science project to advance green chemistry, medical technologies, and photovoltaics.

Gabriele Sadowski, a chemical engineer and TU Dortmund's vice president of research says researchers have also strengthened collaborations across borders and with industry. Physicists from the university are working with colleagues from Russia to improve semiconductor science using the power of electron spin. The university is also part of a joint venture with the pharmaceutical company, Bayer, developing innovative drug-delivery technologies at the chemicals

manufacturing centre of Leverkusen.

The university has a faculty of 2,300 and a 2017 budget of €330 million, including third-party funding. **ANJA KRIEGER**

WAGENINGEN UNIVERSITY & RESEARCH

THE NETHERLANDS | 2015 FC: 34.45* | 2017 FC: 50.55

The founding of Wageningen University in 1876 marked the beginning of the Netherlands' agricultural education programme, a legacy upheld to this day. In 2016, the university merged with its eight research institutes to form one entity: Wageningen University & Research (WUR).

Between 2015 and 2017, Wageningen University's fractional count (FC) increased 62% to 43.59. Adding the newly incorporated institutions to the 2017 figure increased the FC to 50.55. Particularly notable, has been WUR's growth in Earth and environmental sciences, from 11.77 to 18.82.

With a full-time faculty of almost 5,000, the university focusses on three core areas — food production, living environment, and livelihood — covering such diverse areas as non-chemical pest control, a faster route to vaccines, and improving the sustainability of banana plantations. WUR has also prioritized investments in broad research themes, such as resource-use efficiency, metropolitan solutions, synthetic biology, and resilience.

Biogeochemist Dolf Weijers is among its most prolific researchers in the index. Most recently, he led a study into the 700-million-year-long evolution of the hormone auxin, which controls processes involved in plant growth and development. **BN**

HUNAN UNIVERSITY

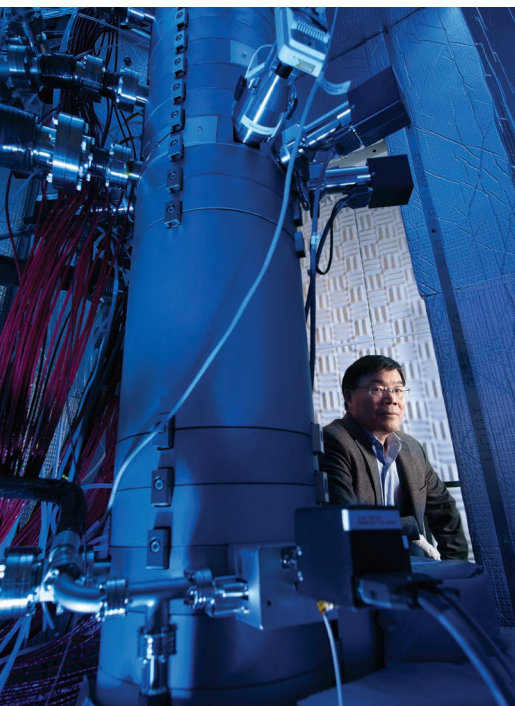
CHINA

PHYSICAL SCIENCES FC 2015: 4.5 | 2017 FC: 23.46

Located in the city of Changsha in south-central China, Hunan University (HNU) has seen remarkable growth in physical sciences, with its output in the field increasing more than four-fold in the index between 2015 and 2017, easing declines in chemistry.

Yet HNU's predominant strengths remain chemistry and materials science. It has two state-funded key labs — China's elite national laboratories — one for vehicle design and manufacturing, and one for biosensing and data-driven chemistry analysis. The latter accounts for roughly half of HNU's research in the index.

Chemist, Xidong Duan, of the biosensing lab says that increased support from the central Chinese government has contributed to HNU's standing, as well as university policies that have sought close collaborations



Xiaoqing Pan, director of the newly established Irvine Materials Research Institute, University of California, Irvine.

with top scientists abroad. In 2017, he and UCLA's Xiangfeng Duan published a paper in *Science* demonstrating a method for producing super-thin semiconductors with intricate structures just a few atoms thick.

Hunan has 1,950 faculty, of which more than 1,400 are professors. **MZ**

MARTIN LUTHER UNIVERSITY HALLE-WITTENBERG

GERMANY | 2015 WFC: 18.81 | 2017 WFC: 29.18

Martin Luther University Halle-Wittenberg was established in 1817 as the largest university in Germany's state of Saxony-Anhalt. The Nazi regime and subsequent politicisation of academia in East Germany took a toll on the university's faculty and programmes. After the fall of the Berlin Wall, MLU emerged as a medium-sized university, with around 340 professors among its faculty in 2017, offering the panoply of science subjects. Its research and teaching budget in 2017 was €205 million (US\$238 million), excluding third-party funding.

MLU's rise in the Nature Index has been driven by its chemistry output: between 2015 and 2017, its contribution to articles in this field almost tripled. To Wolfgang Binder, dean of the faculty of natural sciences, MLU's interdisciplinary approach of connecting physics, chemistry and biology has laid the groundwork for this success.

MLU's Institute of Chemistry conducts research on subjects ranging from nano-structured and self-healing polymers, to the role of protein misfolding in Alzheimer's and Parkinson's diseases, and liquid crystals used in flat-panel displays (LCDs). In 2016, MLU chemist, Carsten Tschierske and his team, in collaboration with Trinity College Dublin, published research on materials that could make liquid-crystal technology faster and more energy-efficient. **AK**

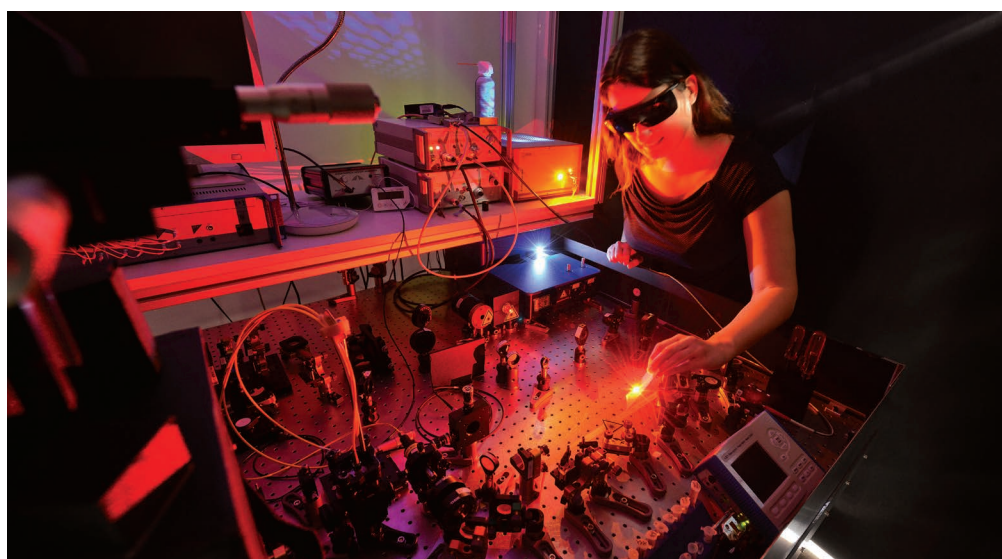
FUDAN UNIVERSITY

CHINA

E&E SCIENCES 2015 FC: 5.67 | 2017 FC: 18.40

Fudan University in Shanghai is a member of the elite C9 League — nine schools sometimes referred to as China's Ivy League. The designation comes directly from the Chinese government, and accords it extra resources and funding. This has been key in luring international researchers, and getting more Chinese researchers who studied overseas to return.

As the government pushes to clean up China's dangerously polluted air and mitigate the effects of climate change, Fudan is helping lead the way. The government's push



Laser devices are used by polymer physicists at Martin Luther University Halle-Wittenberg to determine how proteins are folded.

has given the university many opportunities, says environmental chemist Zhen Ma, whose particular focus is on practical solutions.

Fudan's output in Earth and environmental sciences more than tripled in the index between 2015 and 2017. Work led by Kan Haidong of Fudan's School of Public Health published in 2017 showed that exposure to fine particulate matter less than 2.5 microns in diameter (PM_{2.5}) is linked to blood inflammation and the production of stress hormones, increasing cardiovascular risk.

The university's output in Earth and environmental sciences is set to grow further: in April 2018, it opened a department for atmospheric and oceanic research. **MZ**

UNIVERSITY OF OSLO

NORWAY | 2015 FC: 43.46 | 2017 FC: 65.58

Every five years, the Research Council of Norway offers national research institutions generous ten-year grants to set up centres of excellence in key research areas. The University of Oslo (UiO) has won 17 of the 44 centres since the programme's launch in 2003. Nine centres are active, covering subjects ranging from solar physics to multilingualism and Earth dynamics.

Four centres have been in Earth and environmental sciences, where the university's growth in the index has been particularly strong. UiO's contribution to the authorship in this field has nearly doubled over the past three years, exceeding its 51% growth overall.

The centres give the best and most active researchers more money, says Bjørn Jamtveit, a geophysicist at UiO. The university has also received a significant increase in government funding for climate science, Jamtveit says, which has spurred high-quality

publications and participation in collaborations with universities such as the University of California, Berkeley, UC San Diego, and Ecole Normale Supérieure de Lyon.

Knut Breivik, an environmental scientist at the Norwegian Institute for Air Research, says collaborations between the school and research institutes encourage a mix of basic and applied environmental research.

The research budget at UiO was 3.39 billion NOK (US\$404 million) in 2016, with about 3,635 faculty performing research. **NS**

QUFU NORMAL UNIVERSITY

CHINA | 2015 FC: 1.21 | 2017 FC: 12.66

The city of Qufu has a long history of scholarship: it is the hometown of Confucius.

Qufu Normal University (QFNU) is just coming into its own as a presence in the index. In the last three years, it has seen its fractional count multiply by the largest factor of any institution globally, from 1.21 to 12.66. Its papers in the journals tracked by the index are in chemistry and physics.

Among them, analytic chemist, Fengli Qu, has co-authored a flurry of articles on efficient catalysts for the large-scale production of hydrogen, which has promise as a clean source of fuel and long-term storage reservoir for renewable energy. In the past few years, QFNU has transformed itself from a traditional teaching university to a comprehensive research university, says Qu. One of the most effective policies for improvement, he says, has been the introduction of bonus incentives for research. The university has more than 700 professor positions. In 2002, it expanded to a second campus located in Rizhao, on the Yellow Sea — a city known for its sustainability and adoption of solar water heaters in every new building. **MZ**

MARKUS SCHOLZ/MLU

GREEN SHOOTS

While many of the top institutions have benefited from centuries of steady scientific activity, these younger contenders are reaching for the sky. Established as universities after 1988, they have made significant progress in contributions to 82 high-quality journals over the past three years. We profile six of the leading young institutions.

HOMI BHABHA NATIONAL INSTITUTE

INDIA | FC 2015: 29.84 | FC 2017: 50.20

PHYSICAL SCIENCES 2015 FC: 21.25 | 2017 FC: 33.99

A India's Department of Atomic Energy (DAE) created Homi Bhabha National Institute (HBNI) in 2005 to advance the country's capabilities in nuclear science and engineering. Named after Homi Jehangir Bhabha, considered the father of India's nuclear programme, the university was formed from 10 DAE institutions focusing on nuclear physics and mathematics. Among them is the Variable Energy Cyclotron Centre in Kolkata, which houses India's first major accelerator for peering inside the atomic nucleus. In 2016, an 11th member, the National Institute of Science Education and Research in Bhubhaneswar, became part of the university.

HBNI has increased its contribution to papers in the index by 68.2% since 2015, and is the top-ranked rising star in India in the physical sciences. Papers cover theoretical investigations into subatomic particles, from neutrinos to mesons. Researchers at the institution have also contributed to work on harnessing the Sun's heat to halve the fossil-fuel consumption of existing coal-fired power plants, and the use of molecular assemblies to prevent amyloid build-up in the body, a hallmark of many diseases including Parkinson's. HBNI received 15 million Indian rupees (US\$212,000) in grants and subsidies in 2016–2017, and has 1,042 research faculty. **SMRITI MALLAPATY**

SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

CHINA | FC 2015: 15.7 | FC 2017: 67.6

B Southern University of Science and Technology (SUSTech) is positioned for success in Shenzhen, a city often referred to as China's Silicon Valley. Between 2015 and 2017, funding at SUSTech increased from 153 to 600 million yuan (US\$87.8 million), largely from the Shenzhen government and National Natural Science Foundation of China. With the city's support, SUSTech is also building the country's most advanced cryo-electron microscope laboratory for revealing the structure of biomolecules.

Shenzhen's investments are yielding results. Since 2015, SUSTech has more than quadrupled its contribution to the Nature Index — making it the top rising star under 30, with strengths in chemistry, and rapid growth in the physical sciences. Among its top index contributors is Xumu Zhang, who has a chemical reaction named after him.

Some 90% of the university's 300-strong faculty have overseas experience, with the majority of courses taught in English. SUSTech has enlisted 60 staff members through China's national Thousand Talents Plan, which offers incentives to Chinese professors, trained abroad, to return to China.

The university also has policies that encourage high-quality work, such as a meritocratic appraisal system that makes it easy for researchers to identify promotion opportunities, and deters nepotism. **SARAH O'MEARA**

INDIAN INSTITUTES OF SCIENCE EDUCATION AND RESEARCH

INDIA | FC 2015: 79.01 | FC 2017: 98.21

C Starting with two institutes in 2006 in Pune and Kolkata, there are now seven Indian Institutes of Science Education and Research (IISER), with two set up in 2015 in Berhampur and Tirupati, on the central-eastern coast.

The IISERs were established to bring high-quality science teaching and research under one roof. It was an unusual model for the country. University faculty, burdened with teaching loads, poor funding and limited infrastructure, did not engage in research, while the elite research institutes did not teach. "The IISERs have the right combination of both components," says Soumitro Banerjee, a physicist at IISER, Kolkata.

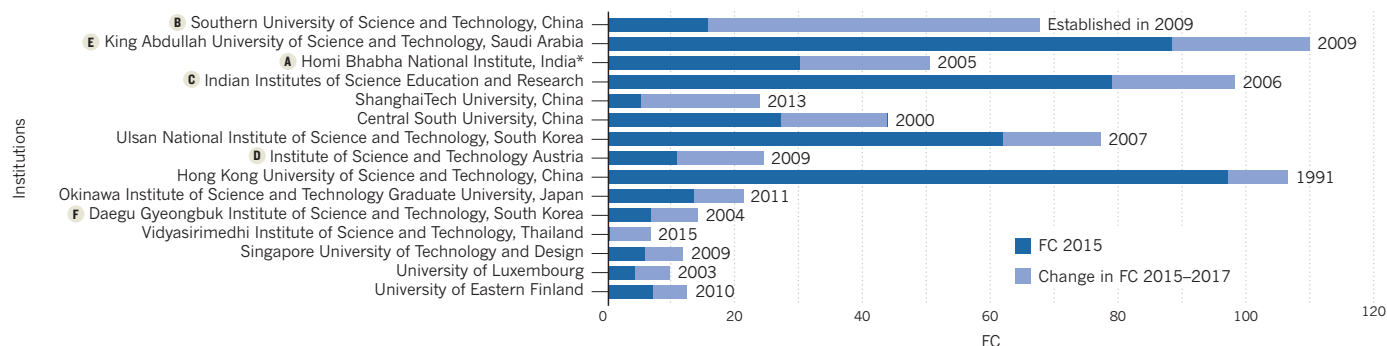
The Indian government now recognizes several IISERs as Institutes of National Importance, a designation given to institutions that play a role in developing highly skilled individuals, which affords them additional funding.

Collectively, the institutes have increased their contribution to science tracked by the index by 24% since 2015.

The mandate to balance teaching with research has engaged undergraduates in research, some publishing results by completion of their master's degree. The IISERs' research budget was 6.3 billion Indian rupees (US\$89 million) in 2016–2017. **T.V. PADMA**

15 UNDER 30

All under 30 years old, these universities are the top 15 among the under thirties in the Nature Index, ranked by their growth in fractional count (FC) 2015–2017. Their success has been achieved largely without the advantage of time, though some have inherited a head start from predecessor institutions.





Matilda Peruzzo, a PhD student in Johannes Fink's group at the Institute of Science and Technology Austria. Her experiments seek to characterize new types of superconducting qubits.

INSTITUTE OF SCIENCE AND TECHNOLOGY AUSTRIA

AUSTRIA | FC 2015: 10.88 | FC 2017: 24.37

D Located in Klosterneuberg, just outside Vienna, researchers at the Institute of Science and Technology Austria (IST Austria) are encouraged to explore their curiosities, far from the fixation on impact. Established by the federal government in 2009 to focus on basic research, “the institute values science for its own sake and not its potential to bring immediate benefit,” says biologist, Fyodor Kondrashov, who runs a research group that uses an array of tools — mathematical modelling, bioinformatics data, and experiments — to study evolution.

The graduate-only university, which is organised into research groups with no departments, encourages students to be collaborative and broadminded, says Kondrashov: students spend the first year rotating among groups to find the best fit for their interests.

Researchers have thrived in the intellectual environment, more than doubling their contribution to the Nature Index since 2015, with the life sciences accounting for two-thirds of this output. Among IST Austria's most prolific teams are those determining how immune cells move and change shape, how plants adapt to changing environments, and how neuronal networks process information.

IST Austria supports 700 staff, and 142 PhD students. The Austrian government committed €290 million (US\$336 million) for its first 10 years, with researchers securing an additional €100 million in third-party funds. **GEMMA CONROY**

KING ABDULLAH UNIVERSITY OF SCIENCE AND TECHNOLOGY

SAUDI ARABIA | FC 2015: 88.42 | FC 2017: 109.98

E King Abdullah University of Science and Technology (KAUST) launched in 2009 to help move Saudi Arabia's oil-dependent economy to one more reliant on knowledge. It supports 710 research scientists in 11 centres tackling clean combustion, water desalination and solar energy, among other transformational technologies.

KAUST is the fastest-growing institution in Western Asia in the Nature Index 2015–2017, increasing its contribution in the life sciences by 79.5% to a fractional count (FC) of 10.5 in 2017, and growing by 57% in the physical sciences to an FC of 47.28.

Its strongest subject, however, is chemistry, bolstered by partnerships with industry giants Saudi Aramco, Saudi Basic Industries Corporation and Dow Saudi Arabia. These partnerships offer researchers start-up funding and job opportunities. In May 2018, KAUST welcomed the Dow Innovation Center, which will facilitate applied research into oil and gas technologies, sustainable construction, and industrial chemicals.

The university's deep pockets for research and salaries have drawn foreign talent. Among them is chemical scientist, Yu Han, who moved to KAUST in 2009, and stands out in the index journals for his work on nanoporous materials used to purify water and clean fuel emissions. Some 95% of KAUST faculty are international, and more than three-quarters of papers co-authored at the university in the journals tracked by the index are with international partners. **GC**

DAEGU GYEONGBUK INSTITUTE OF SCIENCE AND TECHNOLOGY

SOUTH KOREA | FC 2015: 6.73 | FC 2017: 13.99

F Situated in South Korea's manufacturing centre, Daegu Gyeongbuk Institute of Science and Technology (DGIST) was established in 2004 to boost Daegu's economy through innovation. Four years later, DGIST expanded into a university, supporting 234 researchers addressing societal challenges in six key research areas, including emerging materials, green energy and medical robotics.

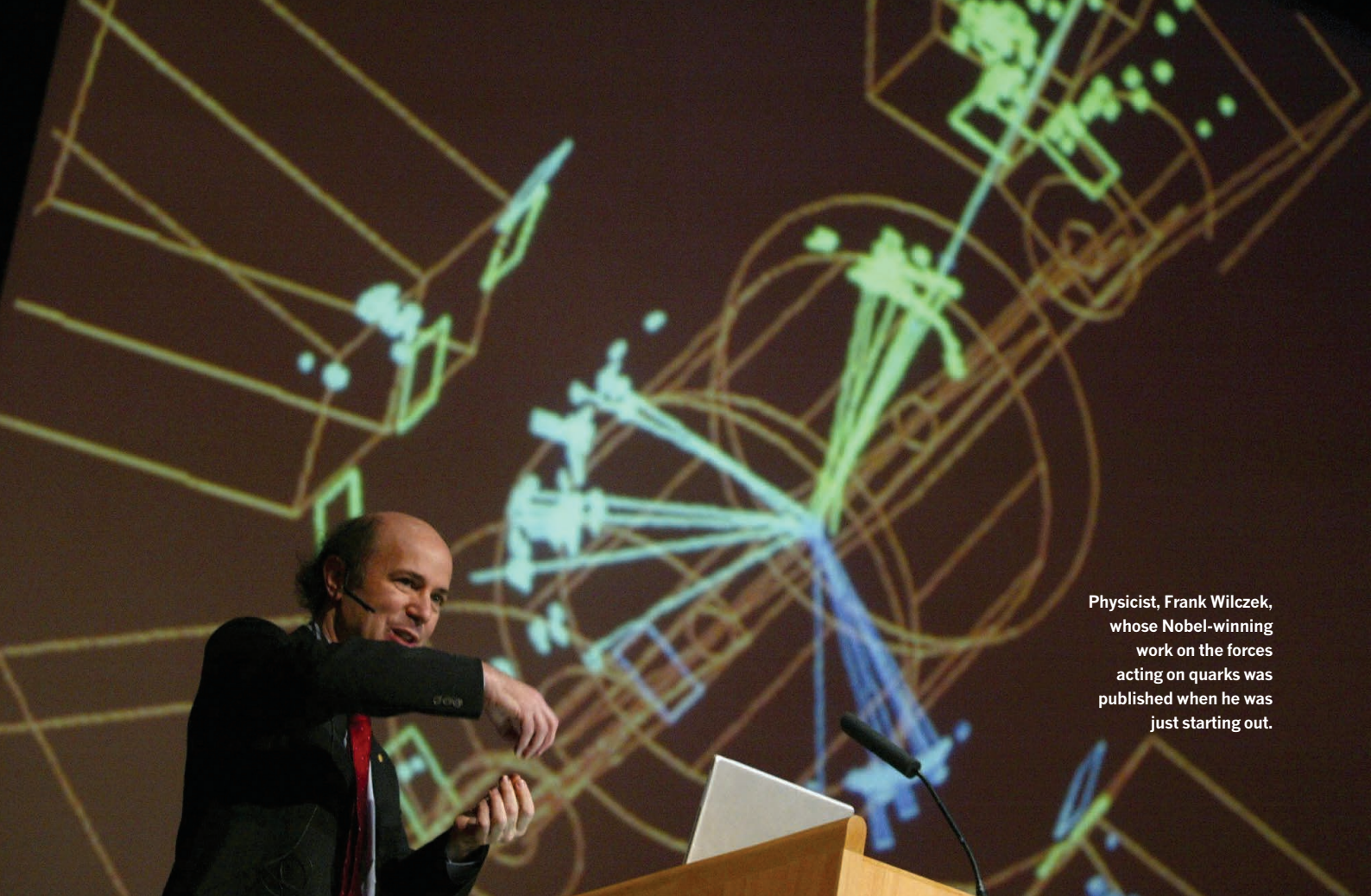
The university encourages commercial partnerships, and has 11 start-ups developing technologies such as rehabilitation exercise devices, robots that work on building maintenance, and a fire-retardant insulation material.

With three other national institutes advancing science across the country, in Daejeon, Gwangju and Ulsan, DGIST president, Sang Hyuk Son, says the university stands out for its interdisciplinarity: roboticists team up with life scientists, while solar energy specialists work with materials scientists.

DGIST's contribution to articles tracked by the index has doubled over the past three years to a fractional count of 13.99 in 2017. Chemistry accounts for more than half of its overall output. In 2017, for example, DGIST researchers developed a high-resolution imaging technique to analyse living biological samples without chemical pre-treatment, which could be applied in medical diagnosis and drug screening. **GC ■**



A researcher at the Convergence Research Center for Wellness, Daegu Gyeongbuk Institute of Science and Technology, in South Korea, demonstrates a robot that can walk and run.



Physicist, Frank Wilczek, whose Nobel-winning work on the forces acting on quarks was published when he was just starting out.

PREDICTING SCIENTIFIC SUCCESS

Even sophisticated, data-driven models of academic careers have trouble forecasting the highs and lows.

SMRITI MALLAPATY

When Frank Wilczek was a graduate student in his early twenties, he published work on the forces holding quarks together that later won him a Nobel Prize.

At the other end of a career span, John Fenn, a retired analytic chemist in his seventies, developed the award-winning technique for analysing large proteins using mass spectrometry.

From early starters to late bloomers, the timing of a researcher's career high is largely dependent on chance. This was the conclusion of a 2016 study, in which researchers developed a mathematical model to describe publication and citation trends based on the records of thousands of people.

Every piece of work is just as likely to be your highest impact paper as the last, says study co-author Dashun Wang at Northwestern

University's Kellogg School of Management in Evanston, Illinois. "To be a successful scientist, you should just keep drawing the lottery and hope for the best."

Sophisticated new models are using vast data sets to help elucidate the process of scientific discovery, and how it will evolve — including at the level of individual careers. As the volume of this information expands, the resulting algorithms and their predictions will improve.

But, in searching for predictable patterns, and a formula for detecting rising research stars, scientists are finding that success is inherently unpredictable, says Daniel Larremore, a computer scientist at the University of Colorado Boulder.

These models are also beginning to reveal the flaws in the research system and point to ways of correcting them. "Through reverse engineering, we can help create a fairer system that nurtures talented people, no matter

their ethnicity, gender or location," says Roberta Sinatra, a network and data scientist at the Central European University in Hungary, and first author of the 2016 study.

BETTING ON THE BEST

Researchers have had limited success in finding quantitative and objective ways of predicting a scientist's future performance based on their past merits.

Earlier efforts typically involved statistical checks of single or collected metrics to see how well they correlate with reality. In 2007, for example, Jorge Hirsch, a physicist at the University of California, San Diego, published a paper on the predictive power of a popular measure he had invented for determining the scientific impact of an individual — the *h*-index. Hirsch observed a correlation between a researcher's current and future *h*-index.

Several years later, a group led by computer

scientist, Daniel Acuna, now at Syracuse University, developed a formula to estimate an individual's future *h*-index based on several variables, including number of articles, publication in prestigious journals and years since first paper. It accounted for 66% of the variability in the *h*-index of some 3,000 neuroscientists five years later. But some scientists argued that the cumulative nature of the *h*-index overstated its predictability.

Now, mathematicians, network scientists, and physicists are bringing new tools to the challenge. They are creating simple models of the rules of human behaviour, in the same way that the Standard Model explains the existence of the Higgs Boson.

These models exploit rich and accessible long-term data generated about scientists and their scholarly endeavours — from publications and citations, to funding sources, collaborators, mobility, institutional affiliation, ethnicity and gender. But a formula for spotting rising research stars is still elusive. In detecting career trends, the models are also revealing predictive limits.

CHANCE DISCOVERY

Those who study the trajectories of scientific careers had long assumed that researchers were at their most creative early in their careers. Sinatra and Wang's 2016 study proved otherwise. They found that a constant and unique value known as *Q*, derived from an individual's long-term citation and publication record, could determine the number of citations that their best paper would achieve, but the timing of that paper was anybody's guess. The higher a researcher's *Q* factor, the higher the impact of their paper.

In a recent study covering a shorter publication window, Wang and Sinatra showed that a career high is typically characterized by a slew of several highly cited papers. "All of an individual's best works tend to happen within that hot streak," says Wang. And while most scientists will experience such a creative burst, it will probably only happen once in their career.

A 2017 study by Larremore also deconstructed the fast-early-peak, slow-slump pattern of productivity. In an analysis of more than 2,000 computer scientists and 200,000 publications, he found that while the researchers' collective publication trajectory followed the rise-fall pattern, it could only explain the productivity of one in every five scientists.

Paper citations don't always follow a reliable pattern either, which makes it difficult to predict career trajectories based on them. Some papers lie dormant for many years before gaining citation traction. A 2015 citation analysis of 22 million articles spanning more than a century found that there are many examples of such 'sleeping beauties'. Among them is a 1955 paper by Eugene Garfield on the utility of a citation index, which caught the research community's attention some half a century later.



First x-ray photograph of a human, in 1895.

While emerging algorithms can potentially anticipate incremental advances in science, such as the observation of gravitational waves, it is beyond their capacity to predict the accidental isolation of penicillin, or the serendipitous discovery of x-rays, as it is beyond the scope of most humans.

"Any kind of model that makes strong bets on the trends of the past is likely to perpetuate the kinds of problems that we have now, without leaving us open to the weird and unexpected innovations that no-one sees coming," says Larremore.

Models of scientists' careers don't need

to be good predictors to be useful, says Vincent Traag, a computational social scientist at the Centre for Science and Technology Studies, Leiden University. By allowing researchers to uncover the mechanisms underlying the phenomena they observe — how science itself works — "we can start thinking of how to address questions such as the replicability crisis, publication biases, and inappropriate incentives," says Traag.

Gaps in the publication records of individuals expose the many lost opportunities — from those who have abandoned academia out of a sense of failure, or to raise children, or for unexplained reasons.

"The big piece of the puzzle that is missing is a quantitative understanding of failure," says Wang, who is analysing grant application data from the US National Institutes of Health to capture signals not just of acceptance, but also rejection. "It happens all the time, yet we know so little about it."

When it comes to tracking talent, some traits have little to do with merit. Studies of the *h*-index, for example, have found that women are cited less than men.

"If we put this into an approach that predicts impact, then it would favour men, rather than women," says Sinatra, who is working on developing data-driven measures and models to identify the source and contribution of forms of bias so they can be corrected, and not perpetuated in predictive modelling.

"So much of the past 'success' has been correlated with looking and sounding, well, like me — white, male, native English speaking, past affiliation with Harvard," says Larremore. "There is a danger of reading too much into the patterns of the past." ■



Algorithms can point to incremental advances, but breakthroughs such as the accidental isolation of penicillin are impossible to predict.

A GUIDE TO THE NATURE INDEX

A description of the terminology and methodology used in this supplement, and a guide to the functionality available free online at natureindex.com

The Nature Index is a database of author affiliations and institutional relationships. The index tracks contributions to research articles published in 82 high-quality natural science journals, chosen by an independent group of researchers.

The Nature Index provides absolute and fractional counts of publication productivity at the institutional and national level and, as such, is an indicator of global high-quality research output and collaboration. Data in the Nature Index are updated regularly, with the most recent 12 months made available under a Creative Commons licence at natureindex.com. The database is compiled by Springer Nature.

NATURE INDEX METRICS

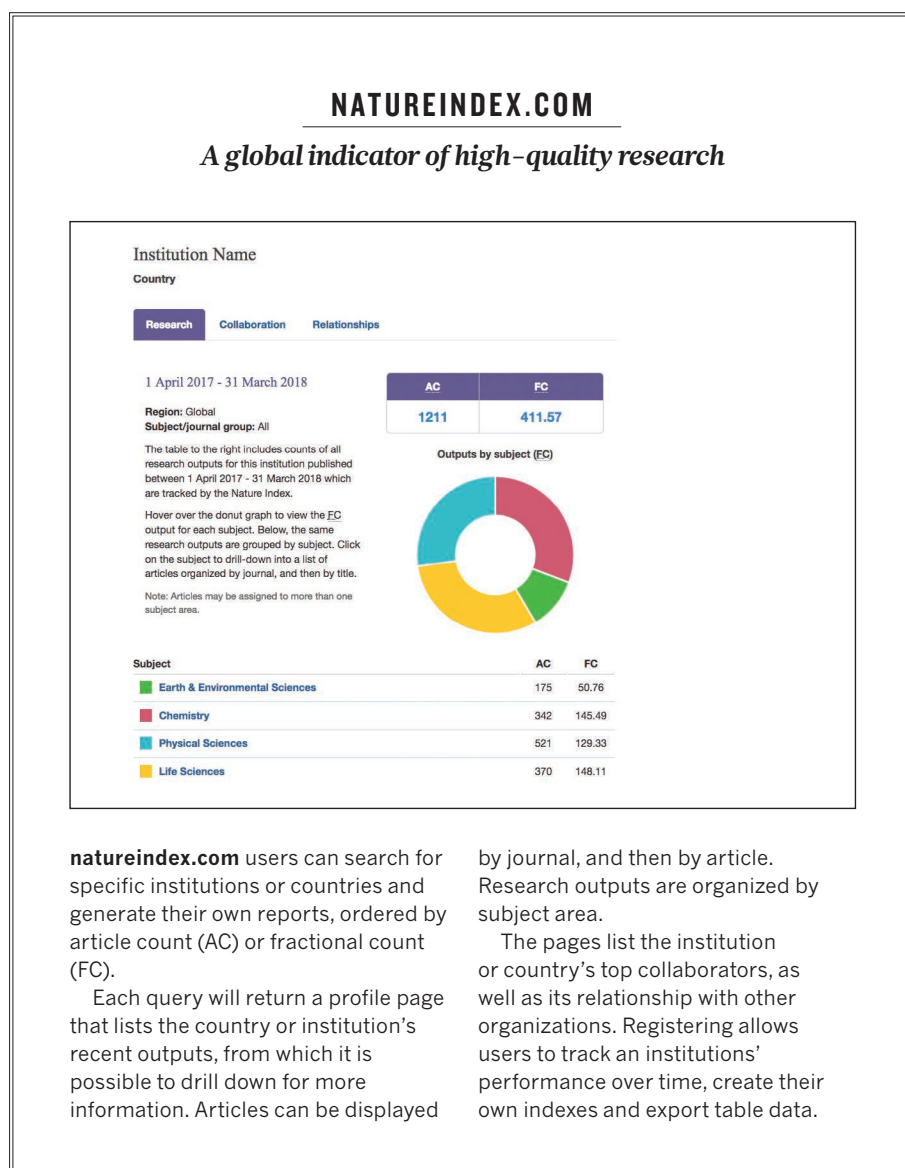
The Nature Index provides several metrics to track research output and collaboration. These include article count, fractional count, and adjusted fractional count.

The simplest is the article count (AC). A country or institution is given an AC of 1 for each article that has at least one author from that country or institution. This is the case regardless of the number of authors an article has, and it means that the same article can contribute to the AC of multiple countries or institutions.

To get a sense of a country's or institution's contribution to an article, and to ensure they are not counted more than once, the Nature Index uses fractional count (FC), which takes into account the share of authorship on each article. The total FC available per article is 1, which is shared among all authors under the assumption that each contributed equally. For instance, an article with 10 authors means that each author receives an FC of 0.1. For authors who are affiliated with more than one institution, the individual author's FC is then split equally between those institutions.

The total FC for an institution is calculated by summing the FC for individual affiliated authors. The process is similar for countries, although complicated by the fact that some institutions have overseas labs that will be counted towards host country totals.

When comparing data over time, FC values are adjusted to 2017 levels to account for the small annual variation in the total number of articles in Nature Index journals. The adjustment of FC values in each year is done



natureindex.com users can search for specific institutions or countries and generate their own reports, ordered by article count (AC) or fractional count (FC).

Each query will return a profile page that lists the country or institution's recent outputs, from which it is possible to drill down for more information. Articles can be displayed

by journal, and then by article. Research outputs are organized by subject area.

The pages list the institution or country's top collaborators, as well as its relationship with other organizations. Registering allows users to track an institutions' performance over time, create their own indexes and export table data.

by calculating the percentage difference in the total number of articles in the index in a given year relative to the number of articles in 2017 and applying this adjustment to FC values.

THE SUPPLEMENT

Nature Index 2018 Rising Stars is based on data from natureindex.com, covering articles published during six years from 1 January 2012 to 31 December 2017 at the country level, and articles from 1 January 2015 to 31 December

2017 at the institution level.

Most analyses within the supplement use adjusted FC as the primary metric. The tables rank institutions by their absolute change in adjusted FC from 2015 to 2017. The tables also provide the percentage change in FC 2015–2017 and an institution's global rank in the 2018 annual tables. Separate tables rank the top academic institutions, top young universities under 30 years old, as well as the top institutions in each subject area. ■